VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV AUTOMATIZACE A MĚŘICÍ TECHNIKY

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF CONTROL AND INSTRUMENTATION

VLIV SELEKCE PŘÍZNAKŮ METODOU HFS NA SHLUKOVOU ANALÝZU

EFFECT OF HFS BASED FEATURE SELECTION ON CLUSTER ANALYSIS

DIPLOMOVÁ PRÁCE
MASTER´S THESIS

AUTOR PRÁCE
AUTHOR

Bc. JAN MALÁSEK

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. PETR HONZÍK, Ph.D.

BRNO 2015

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav automatizace a měřicí techniky

Diplomová práce

magisterský navazující studijní obor
Kybernetika, automatizace a měření

Student: Bc. Jan Malásek ***ID:*** 74870
Ročník: 2 ***Akademický rok:*** 2014/2015

NÁZEV TÉMATU:

Vliv selekce příznaků metodou HFS na shlukovou analýzu

POKYNY PRO VYPRACOVÁNÍ:

Prvním cílem diplomové práce je srovnání různých metod shlukové analýzy se zaměřením na úspěšnost při stanovování počtu shluků a zařazení jednotlivých instancí do správných tříd (testování bude provedeno na datech se známou výstupní třídou). Dalším cílem je implementace metody HFS (hybrid feature selection scheme, Yang et. Al., 2011) určené pro selekci příznaků v úlohách bez učitele a ověření jejího přínosu pro úspěšnost shlukové analýzy.

1. Zpracujte přehled metod používaných pro shlukovou analýzu a pro stanovení počtu shluků.
2. Experimentálně porovnejte úspěšnost metod na datech se známými výstupními třídami z pohledu stanovení počtu shluků a zařazení instancí do správných tříd.
3. Zpracujte stručný přehled metod selekce příznaků v úlohách bez učitele.
4. Podrobněji popište a naprogramujte metodou HFS.
5. Využijte metodu HFS na analýzu naměřených dat a zkuste posoudit její přínos.

DOPORUČENÁ LITERATURA:

Yang, Y., Liao, Y., Meng, G., & Lee, J. A hybrid feature selection scheme for unsupervised learning and its application in bearing fault diagnosis. 2011, Expert Systems With Applications, 38(9), 1311–1320.

Termín zadání: 9.2.2015 ***Termín odevzdání:*** 18.5.2015

Vedoucí práce: Ing. Petr Honzík, Ph.D.
Konzultanti diplomové práce:

doc. Ing. Václav Jirsík, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

Abstrakt

Diplomová práce se zabývá shlukovou analýzou. Shlukování má své základy v mnoha oblastech lidského vědění zahrnujících získávání dat, statistiku, biologii a strojové učení. Hlavní náplní práce je zpracování rešerše metod shlukové analýzy, metod pro stanovení počtu shluků a stručný přehled metod selekce příznaků v úlohách bez učitele. Neméně důležitou součástí je realizace softwaru pro porovnání různých metod shlukové analýzy se zaměřením na úspěšnost při stanovování počtu shluků a řazení jednotlivých instancí do správných tříd. Součástí programu je implementace metody selekce příznaků HFS. Experimentální ověření metod proběhlo ve vývojovém prostředí Matlab.

Ve svém závěru diplomová práce porovnává úspěšnost shlukovacích metod na datech se známými výstupními třídami a posuzuje přínos metody selekce příznaků HFS v úlohách bez učitele pro úspěšnost shlukové analýzy.

Klíčová slova

Shluková analýza, selekce příznaků, učení bez učitele, HFS.

Abstract

Master´s thesis is focused on cluster analysis. Clustering has its roots in many areas, including data mining, statistics, biology and machine learning. The aim of this thesis is to elaborate a recherche of cluster analysis methods, methods for determining number of clusters and a short survey of feature selection methods for unsupervised learning. The very important part of this thesis is software realization for comparing different cluster analysis methods focused on finding optimal number of clusters and sorting data points into correct classes. The program also consists of feature selection HFS method implementation. Experimental methods validation was processed in Matlab environment.

The end of master´s thesis compares success of clustering methods using data with known output classes and assesses contribution of feature selection HFS method for unsupervised learning for quality of cluster analysis.

Keywords

Clustering analysis, feature selection, unsupervised learning, HFS.

Bibliografická citace:

MALÁSEK, J. *Vliv selekce příznaků metodou HFS na shlukovou analýzu*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 110s. Vedoucí diplomové práce Ing. Petr Honzík, Ph.D.

Prohlášení

„Prohlašuji, že svou diplomovou práci na téma Vliv selekce příznaků metodou HFS na shlukovou analýzu jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne: **18. května 2015** ……

podpis autora

Poděkování

Děkuji vedoucímu diplomové práce Ing. Petru Honzíkovi, Ph.D. za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

V Brně dne: **18. května 2015** ……

podpis autora

Obsah

1 ÚVOD

Současný svět je propojen informačními kanály, v nichž se nachází nepředstavitelné množství dat a informací. Příjemce informací musí data roztřídit, analyzovat, optimalizovat je atd. Shluková analýza dat je jednou z nejrozšířenějších technik pro analýzu dat. Využívá se v mnoha vědních oborech pro rozpoznávání vzorů, vyhledávání informací v rozlehlých databázích, zpracování řeči aj.

Diplomová práce zpracovává rešerši metod shlukové analýzy, metod pro stanovení počtu shluků a předkládá stručný přehled metod selekce příznaků v úlohách bez učitele. Dále se zaměřuje na experimentální ověření metod shlukové analýzy a posouzení přínosu metody selekce příznaků HFS v úlohách bez učitele pro úspěšnost shlukové analýzy.

Práce je rozčleněna do tří velkých celků: principy shlukové analýzy, metody shlukové analýzy a experimentální srovnání metod shlukové analýzy.

První část (kap. 2) obsahuje popis základních principů shlukové analýzy. Formuluje její úlohu, popisuje vlastnosti objektů shlukové analýzy, podává přehled kritérií pro stanovení optimálního počtu shluků a stručnou charakteristiku metod selekce příznaků v úlohách bez učitele.

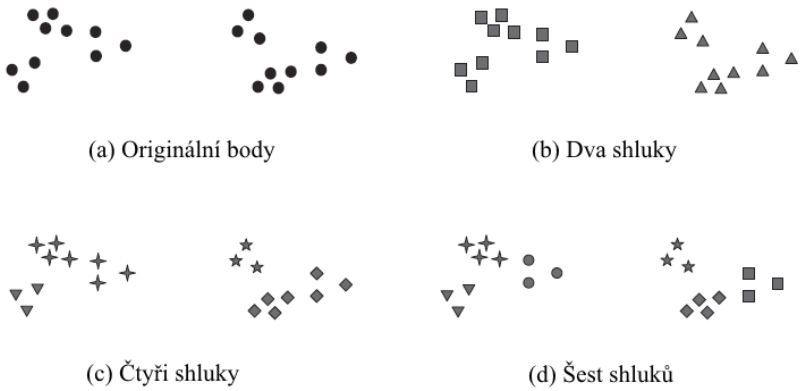
Druhá část (kap. 3) zpracovává přehled metod používaných pro shlukovou analýzu. Je rozdělena do pěti podkapitol, z nichž každá se zabývá jednou z hlavních kategorií metod shlukové analýzy: metody rozkladu, hierarchické metody, metody založené na hustotě, metody založené na mřížce a metody založené na modelu.

Závěrečná třetí část diplomové práce (kap. 4) je zaměřena na experimentální srovnání metod shlukové analýzy se zaměřením na stanovení počtu shluků a zařazení instancí do správných tříd. Pro testování bylo vybráno pět shlukovacích algoritmů. Experiment byl proveden na pěti vybraných datových setech se známými výstupními třídami pro každý shlukovací algoritmus.

Neméně důležitou součástí závěrečné třetí části je ověření přínosu metody HFS určené pro selekci příznaků v úlohách bez učitele pro úspěšnost shlukové analýzy. Obsahuje porovnání úspěšnosti shlukování předzpracovaných dat metodou HFS s výsledky shlukování dat bez selekce příznaků.

2 PRINCIPY SHLUKOVÉ ANALÝZY

Shluková analýza (*cluster analysis*) neboli shlukování (*clustering*), je technika rozdělování množiny datových souborů na podmnožiny. Každá podmnožina představuje shluk objektů, které jsou si navzájem podobné (něčím příbuzné) a současně odlišné (nemající příbuznost) od objektů nacházejících se v jiném shluku. Čím jsou si data v daném shluku podobnější a čím jsou větší rozdíly mezi shluky, tím je shlukování přesnější. Použitím různých metod shlukové analýzy (viz. kap. 3) můžeme ze stejného datového souboru získat různé výstupní shluky, jak ukazuje obr. 2.1. Dvacet bodů je zde rozděleno třemi různými způsoby do shluků. Právě tento obrázek dobře ilustruje, že definice shluku je nepřesná a záleží na povaze zkoumaných dat a požadovaném výsledku.



Obr. 2.1 Shluková analýza [2]

Shlukové analýzy se velmi často využívá v mnoha aplikacích v různých vědních oborech při rozpoznávání znaků (*pattern recognition*), dobývání dat z databází (*data mining*), strojovém učení (*machine learning*), v biologii, medicíně, psychologii aj. [1], [2].

2.1 Úloha shlukové analýzy

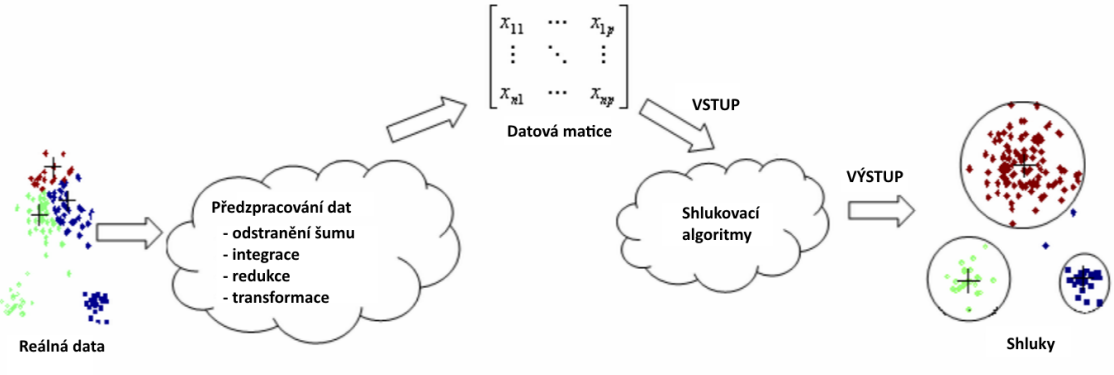
Shlukování je označováno jako jeden ze základních typů získávání informací, a to bez ohledu na skutečnost, zda jsou pro dosažení cíle použity statistické metody, či metody strojového učení. V terminologii těchto metod se rozlišuje

- učení s učitelem (*supervised learning*) a
- učení bez učitele (*unsupervised learning*).

Oba přístupy lze použít ke klasifikaci objektů. Jejím předpokladem je, že máme informace o určitých objektech, které se více či méně navzájem odlišují. Může existovat několik skupin těchto objektů. Cílem je zařadit některé z nich nebo všechny do skupin.

V případě učení s učitelem obsahuje vstupní datový soubor informace o příslušnosti objektů do známých skupin. Cílem je vytvořit model, na základě něhož by mohly být objekty bez známé příslušnosti zařazovány do daných skupin.

Při učení bez učitele není předem známa příslušnost žádného z objektů a obvykle není znám ani počet skupin. Cílem je klasifikovat všechny objekty zahrnuté do analýzy. Tento postup je označován jako shlukování. [4]



Obr. 2.2 Proces shlukové analýzy [3]

Obrázek 2.2 zobrazuje celý proces shlukování. Původní (tzv. surová) data jsou předzpracována různými metodami používanými pro předzpracování dat, jako jsou např.: čištění dat, integrace, transformace nebo redukce dat. Takto připravená data jsou vstupem pro shlukovou analýzu, jejímž výstupem jsou shluky původních dat. [3]

Aby bylo možné dosáhnout výsledných shluků, je potřeba vyřešit celou řadu dílčích úkolů. Prvním problémem je stanovení podobnosti dvou objektů. Aby mohla být podobnost změřena, musí být každý objekt charakterizován pomocí svých vlastností. Například textový dokument je charakterizován klíčovými slovy, minerální voda koncentracemi určitých iontů, rostlina tvarem listu atd.

Shlukování založené na měření podobnosti se nazývá konvenční. Označíme si dva objekty jako A a B. Symbolicky pak můžeme zapsat, že

$$podobnost(A,B) = f\big(vlastnosti(A), vlastnosti(B)\big),$$

tedy podobnost dvou objektů je funkcí jejich vlastností. Interpretace výsledných skupin může být ale v některých případech velmi obtížná díky značnému zjednodušení reality.

Kromě konvenčního shlukování se také používá shlukování konceptuální. Vytvářené shluky jsou založeny na konceptuální soudržnosti, která je funkcí jednak vlastnosti objektů, jednak popisného jazyka L a okolí E. Popisný jazyk je způsob, jakým jsou popsány třídy (skupiny) objektů, a okolí je množina sousedících vzorů. Symbolicky můžeme zapsat, že

$$konceptuální\ soudržnost(A,B) = f(vlastnosti(A), vlastnosti(B), L, E).$$

V praxi se konceptuální shlukování využívá například při analýze textových databází.

2.2 Formulace úlohy shlukové analýzy

Nechť X značí množinu n objektů. Rozklad $\Omega = \{C_1, C_2, \dots, C_m\}$ množiny X je množina disjunktních, neprázdných podmnožin X, které pro $i \neq j$ dohromady tvoří X.

$$C_i \cap C_j = \emptyset, \qquad C_1 \cup C_2 \cup \dots \cup C_m = X.$$

Každá množina C_i se nazývá komponentou rozkladu.

Shlukování je rozklad množiny X. Komponenty tohoto rozkladu se nazývají shluky. Shlukování patří ke třídě úloh následujícího tvaru:

Nechť x je náhodný objekt z množiny objektů X s rozdělením pravděpodobnosti $p_x X \rightarrow \mathbb{R}$**.** Nechť D je množina rozhodnutí taková, že pro každý objekt $x \in X$ seurčí jisté rozhodnutí $d \in D$. Nechť $W: X \times D \rightarrow \mathbb{R}$ je pokutová funkce, jejíž hodnota $W(x, d)$ představuje ztráty v případě volby rozhodnutí d pro objekt x. Při zvoleném rozhodnutí d zde pokuta závisí na známém porovnání x a ne na nepozorovatelném stavu. Pojem nepozorovatelný stav se zde vůbec nevyskytuje.

Cílem úlohy je zkonstruovat strategii $Q: X \rightarrow D$, která minimalizuje hodnotu $\sum_{x \in X} p(x) \cdot W(x, Q(x))$. [8]

2.3 Objekty a znaky

Shluková analýza je mnohem komplexnější proces než samotné přiřazení objektů do shluků, popř. vytvoření hierarchie rozkladů. Součástí analýzy bývá obvykle posouzení, zda mají být do analýzy zahrnuty všechny proměnné nebo jen některé a zhodnocení jejich významu pro analýzu. Většinou je nutné provést i jejich transformaci (standardizaci). Dále je nutné identifikovat odlehlé objekty, které mohou být definovány různými způsoby. Souhrnně můžeme tyto procesy nazvat jako předzpracování dat. Další důležitou součástí shlukové analýzy je stanovení vhodného (optimálního) počtu shluků (viz kapitola 2.4). [5]

Základním přístupem shlukové analýzy je jednoznačné zařazení každého objektu do právě jednoho shluku. Pokud bychom vytvořili tabulku, v níž by řádky představovaly jednotlivé objekty a sloupce jednotlivé shluky, pak by tabulka obsahovala pouze hodnoty 1 (objekt je zařazen do shluku) a 0 (objekt zařazen do shluku není). Přitom hodnota 1 by se v každém řádku vyskytovala pouze jedenkrát. Toto shlukování se nazývá pevné (objekt je buď zařazen, nebo nezařazen) a disjunktní (objekt je zařazen právě do jednoho shluku).

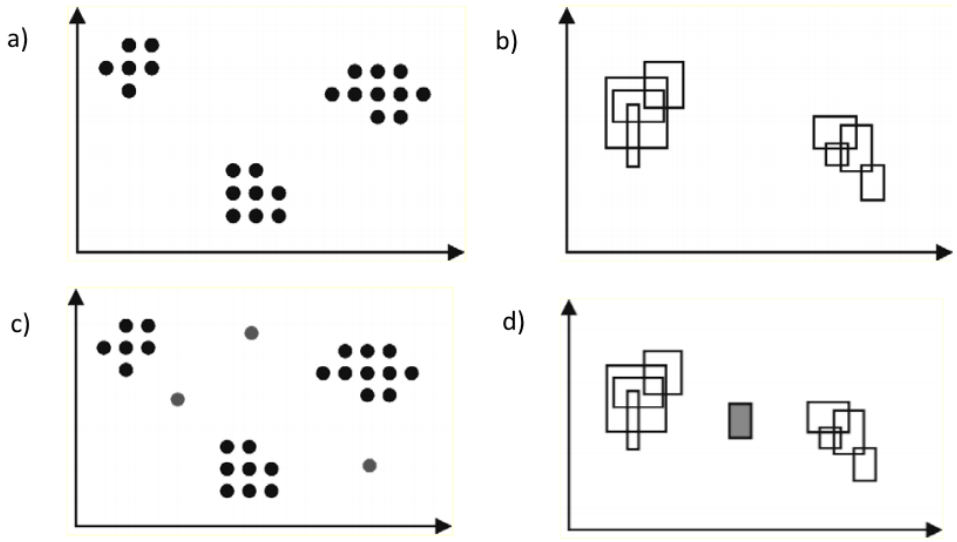
Struktura reálných datových souborů však nebývá takto jednoznačná. Představme si, že potřebujeme stanovit skupiny podobných dokumentů. Pro popis dokumentu vytvoříme seznam slov, podle nichž má být zjišťována podobnost. Dokument pak popsán posloupností hodnot 0 (slovo se v dokumentu nevyskytuje) a 1 (slovo se v dokumentu vyskytuje). Některý dokument může pojednávat jak o aplikaci statistických metod, tak o aplikaci neuronových sítí. Pokud by výsledné shluky

představovaly tematicky zaměřené dokumenty, měl by být zmíněný dokument zařazen do dvou shluků. Výsledkem jsou tedy navzájem překrývající se shluky. Takové shlukování nazýváme jako překrývající se.

Analýza však může jít ještě hlouběji, kdy výsledná tabulka přiřazení objektů do shluků bude místo diskrétních hodnot 0 a 1 obsahovat reálná čísla z intervalu <0; 1>, vyjadřující stupeň příslušnosti objektu k danému shluku. Součet těchto hodnot pro každý jednotlivý objekt se přitom rovná hodnotě jedna. Tehdy hovoříme o fuzzy shlukové analýze. [4]

Na základě typu dat a typu shlukování můžeme vymezit čtyři situace:

1. pevná data a disjunktní shlukování, viz obrázek 2.3a (můžeme rozlišit tři shluky bodů),
2. fuzzy data a disjunktní shlukování, viz obrázek 2.3b (lze rozlišit dva shluky obdélníků),
3. pevná data a překrývající se shlukování, viz obrázek 2.3c (šedě zakreslené body mohou být přiřazeny současně ke dvěma shlukům),
4. fuzzy data a překrývající se shlukování, viz obrázek 2.3d (šedě zakreslený obdélník může být přiřazen současně ke dvěma shlukům).



Obr. 2.3 Grafická reprezentace informačních situací [4]

Jak bylo zmíněno v předcházejícím textu, existují různé typy shlukování, a tudíž i různé typy metod shlukové analýzy, které jsou podrobněji popsány v kapitole 3.

2.3.1 Typy znaků

Při shlukování je každý objekt reprezentován množinou vlastností. V praxi se stanoví znaky (proměnné veličiny), které je třeba sledovat. Tyto proměnné mohou být různých typů, například kvantitativní, jako délka a šířka řeky, počet zvířat v zoo apod.

Jiným typem jsou proměnné kvalitativní, které nenesou číselnou informaci. Ty dělíme na *ordinální*, jejichž hodnoty lze uspořádat, nemusí se však jednat o čísla (intenzita svitu lampy může být malá, střední, silná, případně žádná). Obdobně lze slovně charakterizovat velikost zvířat (malé, střední, velké). Dalším typem kvalitativních proměnných je proměnná nominální, jejíž hodnoty uspořádat nelze. Budeme-li charakterizovat rostlinu pomocí listů, pak můžeme rozlišit listy celistvé, složené, případně zvláštní listové útvary. Speciálním typem je proměnná dichotomická, která nabývá pouze dvou různých hodnot (ano x ne, 1 x 0, kuřák x nekuřák apod.). Lze využít i popis objektů pomocí znaků, které označujeme jako symbolické (interval hodnot). Pokud nemáme k dispozici jednu konkrétní hodnotu, označují se taková data jako neurčitá nebo mlhavá (*fuzzy*). [6]

2.3.2 Identifikace odlehlých objektů

Shluková analýza je velmi citlivá na přítomnost nevýznamných znaků. Výsledek shlukování je dále závislý na přítomnost odlehlých objektů, které se silně odlišují od všech ostatních objektů. Při použití základních shlukovacích algoritmů tvoří takové objekty samostatné shluky. Odlehlé objekty mohou představovat buď skutečně odchýlené objekty, které nejsou představiteli analyzované populace, nebo chybný výběr objektu z populace, který způsobí nevhodné zastoupení původní populace. [6]

2.3.3 Chybějící objekty

V reálných datových souborech obvykle chybí některé údaje. Je možných několik příčin, jako např.: nezjištění příslušné hodnoty, získání nesmyslné nebo nepravděpodobné hodnoty, případně chyba při vstupu dat. [4]

Při shlukové analýze existují tři základní postupy:

- nahrazení chybějící hodnoty,
- vynechání objektu, u něhož některý údaj chybí,
- použití speciální míry pro zjištění meziobjektové podobnosti (viz kapitola 2.3.4).

2.3.4 Míry podobnosti a vzdálenosti

V různých etapách většiny základních algoritmů shlukování posuzujeme podobnost dvou objektů (resp. proměnných či kategorií), popř. podobnost objektu (proměnné, kategorie) a shluku a podobnost dvou shluků. Tyto podobnosti se dále

využívají při hodnocení vytvořených shluků, ať už z důvodu výběru rozkladu, pokud bylo aplikováno více metod, nebo z důvodu stanovení vhodného počtu shluků.

Ve shlukové analýze se obvykle podobnost posuzuje zprostředkovaně pomocí odlišností. Čím méně jsou objekty odlišné, tím jsou si podobnější. Každou míru podobnosti lze převést na míru nepodobnosti a naopak. Uvažujeme-li míru podobnosti $S \in \langle 0; 1 \rangle$, pak se jako transformace na míru podobnosti $D \geq 0$ používá odečtení od jedničky, tj. $D = 1 - S$. Obsahuje-li datová matice hodnoty kvantitativních proměnných, pak se mezi vektory, které charakterizují objekty, počítá vzdálenost. Pokud by objekty byly popsány pomocí dvou proměnných, lze je znázornit jako body ve dvourozměrném prostoru a spočítat jejich vzdálenost. Obdobně se postupuje i v případě vektorů větších rozměrů. Pro výpočet vzdálenosti mezi objekty se v praxi nejčastěji používají následující metriky, které lze obyčejně zařadit do jedné ze tří základních skupin:

a) Korelační míry

Základní mírou podobnosti dvou objektů či znaků x_i a x_j, vyjádřených v kardinální škále může být Pearsonův párový korelační koeficient r. Objekty jsou si tím podobnější, čím je jejich párový koeficient větší a bližší jedničce. V případě ordinální škály (pořadová čísla) je analogickou měrou podobnosti Spearmanův korelační koeficient. Obyčejně se vychází z transponované matice dat X^T, kdy sloupce představují objekty a řádky pak znaky. Korelační koeficienty mezi dvěma sloupci matice X^T představují korelaci mezi dvojici objektů. Tomu odpovídá podobnost jejich profilů v profilovém diagramu. Vysoká korelace prozrazuje vysokou „podobnost“ a nízká korelace pak „nepodobnost“ profilů. [6]

b) Míry vzdálenosti

Představují nejčastěji užívané míry, založené na prezentaci objektů v prostoru, jehož souřadnice tvoří jednotlivé znaky. Nejčastější vzdálenostní mírou je euklidovská vzdálenost zvaná také geometrická metrika, která představuje délku přepony pravoúhlého trojúhelníka a její výpočet je založen na Pythagorově větě. Platí, že vzdálenost

$$d_E(x_k, x_i) = \sqrt{\sum_{j=1}^{p} \left(x_{kj} - x_{ij}\right)^2} = \|x_k - x_i\| \tag{2.1}$$

představuje standardní typ vzdálenosti. K dalším možným způsobům vyjádření vztahu obou objektů patří čtvercová euklidovská vzdálenost, která tvoří základ Wardovy metody shlukování (viz kap. 3.2.1) a je vyjádřena vzorcem

$$d_{ES}(x_k, x_i) = \sum_{j=1}^{p} \left(x_{kj} - x_{ij}\right)^2. \tag{2.2}$$

Často je užívaná *Manhattanská vzdálenost* zvaná také vzdálenost městských bloků (*city-block*) nebo *Hammingova metrika*, definovaná vztahem

$$d_{CB}(x_k, x_i) = \sum_{j=1}^{p} |x_{kj} - x_{ij}|. \tag{2.3}$$

Před použitím této metody se musíme ujistit, že znaky spolu nekorelují. Když tato podmínka není splněna, shluky jsou nesprávné. *Čebyševovu vzdálenost* vyjádříme ve tvaru

$$d_C(x_k, x_i) = max_j |x_{kj} - x_{ij}|. \tag{2.4}$$

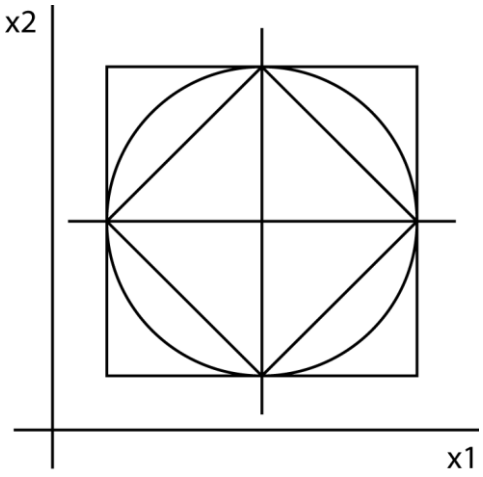
Lanceyova-Williamsova vzdálenost (*Canberra*):

$$d_{LW}(x_k, x_i) = \sum_{j=1}^{p} \frac{|x_{kj} - x_{ij}|}{|x_{kj}| + |x_{ij}|}. \tag{2.5}$$

Obecným vyjádřením vzorců (2.1), (2.3) a (2.4) je *Minkowského metrika*

$$d_M^{(m)}(x_k, x_i) = \sqrt[m]{\sum_{j=1}^{p} |x_{kj} - x_{ij}|^m}, \tag{2.6}$$

kde $m \geq 1$ a čím je větší, tím více je zdůrazňován rozdíl mezi vzdálenými objekty. Tato metrika v sobě zahrnuje uvedené míry jako speciální případy, přičemž $d_C = \lim_{m\to\infty} d_M^{(m)}$. Grafické znázornění výše uvedených vzdáleností ve dvourozměrném prostoru je uvedeno na obrázku 2.4. Pro danou míru mají objekty na obvodu příslušného obrazce shodnou vzdálenost od jeho středu – vnitřní čtverec charakterizuje manhattanskou vzdálenost, kruh euklidovskou vzdálenost a vnější čtverec Čebyševovu vzdálenost.



Obr. 2.4 Grafické znázornění měr vzdálenosti [5]

Všechny dosud uvedené matriky neuvažují závislost mezi znaky. Použijeme-li pro váženou euklidovskou vzdálenost jako čtverec vah inverzní kovariantní matici S^{-1},

odstraníme kromě závislosti na měřících jednotkách také nevýhodu plynoucí z nadměrného vlivu korelovaných proměnných. Výsledkem je *Mahalanobisova vzdálenost* ve tvaru

$$d_{MH}(x_k, x_i) = (x_k - x_i)^T S^{-1}(x_k - x_i). \quad (2.7)$$

Jde vlastně o vzdálenost bodů v prostoru, jehož osy nemusí být ortogonální. Vysoce korelovaný výběr znaků může skrytě převážit celý soubor znaků shlukování. [5], [6]

c) Míry asociace

Míry asociace podobnosti se používají k porovnání objektů, pokud jsou jejich znaky nemetrického charakteru, např. binární proměnné. Je třeba rozlišovat symetrické a asymetrické proměnné. Uveďme příklad, kdy respondent odpověděl na řadu otázek odpovědí *ano* nebo *ne*. Míra asociace pak vyjadřuje stupeň souhlasu každého páru respondentů. Nejjednodušší mírou asociace bude procento souhlasu, kdy oba respondenti odpověděli *ano* nebo *ne*, tedy 1 nebo 0. Rozšíření tohoto jednoduchého „souhlasného koeficientu“ je podstatou míry asociace k vyhodnocování více kategorií nominálních nebo ordinálních znaků.

Například: předpokládejme, že sledujeme asociaci mezi dvěma objekty O_i a O_j. Možné binární odezvy typu 0/1 je pak možno zapsat do tzv. kontingenční tabulky:

		Objekt O_i	
		1	0
Objekt O_j	1	a	b
	0	c	d

Tab. 2.1 Kontingenční tabulka měr asociace [5]

V tabulce 2.1 jsou shrnuty všechny možné kombinace počtu znaků pro dva objekty:

- *a* značí počet znaků, kde mají oba objekty O_i a O_j hodnotu 1 a jde o tzv. pozitivní shodu,
- *b* značí počet znaků, kde má objekt O_j hodnotu 1 a objekt O_i hodnotu 0,
- *c* značí počet znaků, kde má objekt O_j hodnotu 0 a objekt O_i hodnotu 1,
- *d* značí počet znaků, kde mají oba objekty O_i a O_j a hodnotu 0 a jde o tzv. negativní shodu.

Sledujeme, kolikrát pro všechny příznaky objektů O_i a O_j nastaly případy shody a neshody. Na základě počtu zjištěných shod a neshod definujeme různé koeficienty asociace. Pro měření podobnosti mezi dvěma symetrickými proměnnými se používá Sokalův-Michenerův koeficient prosté shody, viz vzorec (2.8). Nepodobnost mezi dvěma asymetrickými proměnnými může být měřena pomocí koeficientů popsaných vzorci (2.9) až (2.15). [5], [6], [7]

a) Sokalův-Michenerův koeficient prosté shody

$$S_{SM} = \frac{a+d}{a+b+c+d} \tag{2.8}$$

b) Russelův-Raoův koeficient

$$S_{RR} = \frac{a}{a+b+c+d}. \tag{2.9}$$

c) Jaccardův (Tanimotův) koeficient podobnosti

$$S_J = \frac{a}{a+b+c} \tag{2.10}$$

d) Hammanův koeficient

$$S_H = \frac{a+d-b-c}{a+b+c+d} \tag{2.11}$$

Hammanův koeficient nabývá hodnot z intervalu <-1,1>, přičemž hodnoty -1 nabývá, pokud se příznaky neshodují ani jednou, 0 nabývá, když je počet shod a neshod v rovnováze a hodnoty +1 v případě úplné shody mezi všemi příznaky.

e) Diceův koeficient (Czekanowského, Sørensonův)

$$S_D = \frac{2a}{2a+b+c} = \frac{2a}{(a+b)+(a+c)} \tag{2.12}$$

f) Korelační koeficient

$$r_B = \frac{ad-bc}{\sqrt{(a+b)+(c+d)(a+c)(b+d)}} \tag{2.13}$$

g) Rogersův-Tanimotův koeficient

$$S_{RT} = \frac{a+d}{a+d+2(b+c)} = \frac{a+d}{(b+c)+(a+b+c+d)} \tag{2.14}$$

h) Ochiaiův koeficient (Kosinova míra pro dvě binární proměnné)

$$S_O = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}} \tag{2.15}$$

2.3.5 Standardizace dat

Většina měr vzdáleností je velmi citlivá na měřítka (stupnice), což vede k různé numerické velikosti znaků. Určité znaky se tak jeví jako dominující a jiné znaky jen málo ovlivňující průběh shlukování. Obecně platí, že znaky s větší mírou proměnlivosti čili větší směrodatnou odchylkou mají menší vliv na míru podobnosti. Je proto výhodné upravit data tak, aby všechny znaky byly souměřitelné. Jedním ze způsobů, jak toho docílit, je standardizace dat.

Standardizace znaků

Nejpoužívanější formou standardizace je normalizace každého znaku do svého Z-skóre, tj. odečtení průměru a dělení směrodatnou odchylkou. Tato standardizace je známa pod

názvem normovací Z-funkce. Tato transformace eliminuje rozdíly v měřítku, mnohdy i řádově se lišících znaků.

Nechť je dána matice dat $\boldsymbol{Z} = (z_{ij})$ typu $n \; x \; p$, jejíž řádky jsou p-rozměrné vektory čísel charakterizující *n* objektů. Standardizaci dat provedeme ve dvou krocích:

1. Vypočteme střední hodnotu $\bar{z}_j$ j-tého znaku z_j a směrodatnou odchylku s_j pro j = 1, 2, … p podle vzorců:

$$\overline{z_j} = \frac{1}{n}\sum_{i=1}^{n} z_{ij} \; , \qquad s_j = \left[\frac{1}{n}\sum_{i=1}^{n}\left(z_{ij} - \bar{z}_j\right)^2\right]^{\frac{1}{2}'} \tag{2.16}$$

2. Původní hodnoty z_{ij} j-tého znaku i-tého objektu přepočteme na tzv. standardizované hodnoty:

$$x_{ij} = \frac{z_{ij}}{\bar{z}_j} \tag{2.17}$$

Tyto standardizované hodnoty znaků mají nyní střední hodnotu rovnu 0 a rozptyl 1.

Výhody standardizace znaků:

- Znaky lze v jednotném měřítku (kde je střední hodnota 0 a rozptyl 1) vzájemně porovnávat snadněji. Kladné hodnoty jsou nad průměrem a záporné hodnoty jsou pod průměrem.
- Se změnou měřítka nedojde k rozdílu mezi standardizovanými znaky.

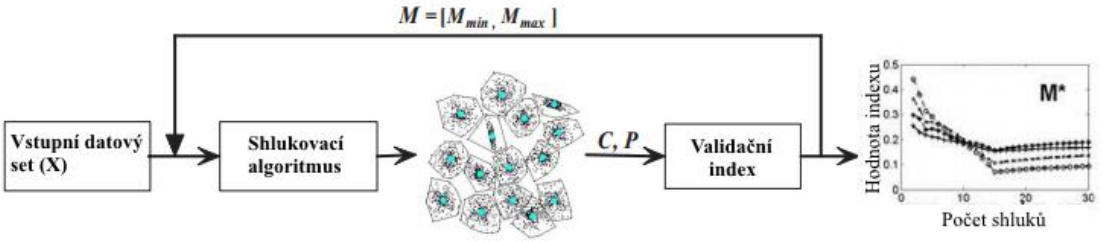
Standardizace objektů

Chceme-li identifikovat shluky dle vzdálenosti, pak standardizace není vhodná. Standardizace objektů neboli řádková standardizace může být však efektivní ve speciálních případech. [6], [8]

2.4 Stanovení optimálního počtu shluků

Stanovení optimálního počtu shluků v datovém souboru je nezbytné pro efektivní a účelné shlukování dat. Mnoho shlukovacích algoritmů je limitováno nutností předem určit a nastavit počet shluků. Pro některé aplikace je možné počet shluků uživatelsky nastavit pomocí expertízy nebo při znalosti daného oboru. V mnoha případech je ale počet shluků předem neznámý. Správné počáteční nastavení tohoto parametru významně ovlivňuje výkonnost algoritmů. Pokud dojde k nevhodnému zvolení počátečních podmínek, může dojít k vytváření nepřesných shluků, jako např. u shlukovacích algoritmů založených na metodě *k*-průměrů.

Pro určení počtu shluků se vyžívá tzv. validačních indexů. Chceme-li určit počet shluků *M**, je parametr *M* optimalizován validačními indexy a ostatní parametry jsou neměnné. Proces určení počtu shluků je zobrazen na obrázku 2.5. Na vstupu je datový set *X*, je zvolen určitý shlukovací algoritmus a určen rozsah [M_{min}, M_{max}], ve kterém se bude iterativně měnit počet shluků. Po každé iteraci je získán výsledek shlukování a spočten validační index. Na základě těchto vyspočtených hodnot se určí nejlepší výsledek shlukování (počet shluků). [33]



Obr. 2.5 Proces validace shluku [33]

Pro vyhodnocení a určení optimálního shlukovacího algoritmu se využívají dvě kritéria:

- **Kompaktnost** (*Compactness*) – prvek každého shluku by měl být co možná nejblíže k dalšímu prvku. Kompaktnost se běžně určuje výpočtem odchylky.
- **Oddělitelnost** (*Separability*) – uvádí, jak moc jsou dva shluky rozdílné. Vypočítává vzdálenost mezi dvěma rozdílnými shluky.

Existují tři přístupy k určení validace shluků:

a) **Externí kritéria** – jsou založena na předchozí znalosti dat. To znamená, že se výsledek shlukování ohodnotí na základě předem specifikované struktury uložené v datovém souboru. Tato externí informace tedy není obsažena v datovém souboru.
b) **Interní kritéria** – jsou založena na informacích, které jsou obsažena v datovém souboru. Interní kritéria se někdy dělí na dvě skupiny: jednu skupinu, která ohodnocuje správnost vztahu mezi daty a očekávaným výsledkem a druhou, jež se soustředí na výslednou stabilitu řešení.
c) **Relativní kritéria** – jsou založena na porovnávání výsledků shlukování jejich srovnáním s výsledky jiných shlukovacích algoritmů. Cílem relativních kritérií je výběr nejlepších shlukovacích algoritmů.

Interní i externí kritéria jsou založena na statistických metodách a mají velké výpočetní nároky. Určení správného počtu shluků pomocí algoritmů není zdaleka jednoznačné, neboť výpočet není jednoduchý. Nalezení správného počtu shluků obvykle závisí na rozložení a měřítku datového souboru, a stejně tak na požadovaném

rozkladu souboru. Pro zjištění počtu shluků existuje velké množství interních i externích kritérií, dále je uveden přehled nejpoužívanějších z nich. [1], [16], [23], [24]

2.4.1 Externí validační kritéria

Externí validační kritéria se využívají pro porovnání výsledků shlukování při znalosti referenčních tříd klasifikovaných objektů.

CA index

Klasifikační kritérium CA (*Classification Accuracy*) porovnává počet správně klasifikovaných datových bodů ve výsledku shlukování se známými třídami objektů. K výpočtu hodnoty kritéria CA je nutné, aby byl každý výsledný shluk přeznačen většinovým (*majority*) štítkem, který označuje shluk, ze kterého pochází nejvíce datových bodů. Jestliže m_i je počet datových bodů s většinovým štítkem (názvem třídy) ve shluku i, poté můžeme CA kritérium definovat jako poměr správně klasifikovaných datových bodů k celkovému počtu datových bodů N v datovém souboru:

$$CA = \frac{\sum_{i=1}^{K}(m_i)}{N}. \tag{2.18}$$

Hodnota kritéria CA se mění v rozsahu <0; 1>. Jestliže je hodnota kritéria CA rovna 1, značí to, že všechna data se nachází ve správných shlucích a každý shluk obsahuje pouze data se stejným štítkem.

Rand Index

Validační kritérium RI (*Rand Index*) bere v potaz počet párových objektů, které existují ve stejných a odlišných shlucích. RI definujeme jako:

$$RI = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}, \tag{2.19}$$

kde n_{11} představuje počet párových objektů, jež jsou ve stejném shluku v obou porovnávaných datových celcích, n_{00} jsou párové objekty umístěné v rozdílných shlucích v obou celcích, n_{10} označuje počet párů, jež se nachází ve stejném shluku v prvním celku, ale v jiném shluku ve druhém celku a n_{01} je počet párových objektů v jiném shluku prvního celku, ale nacházejících se ve stejném shluku ve druhém celku. Lze tedy říci, že n_{11} a n_{00} označují shodu mezi oběma datovými celky, zatímco n_{10} a n_{01} udávají počet neshod, tedy špatných přiřazení bodů do shluků.

Kritérium RI nabývá hodnot v rozsahu <0; 1>. Čím víc se hodnota kritéria blíží jedničce, tím je větší shoda mezi porovnávanými datovými celky.

AR index

Hlavním nedostatkem výše zmíněného Rand indexu je, že při porovnávání různých segmentů neočekává jako výslednou hodnotu nulu. Tento nedostatek byl

odstraněn s příchodem tzv. přizpůsobeného (*adjusted*) *Rand indexu* (*AR*). AR index mezi dvěma segmenty je definován jako:

$$AR = \frac{n_{11} - \frac{(n_{11}+n_{10})(n_{11}+n_{01})}{n_{00}}}{\frac{(n_{11}+n_{10})(n_{11}+n_{01})}{2} - \frac{(n_{11}+n_{10})(n_{11}+n_{01})}{n_{00}}}. \quad (2.20)$$

Stejně jako u Rand indexu platí, že čím vyšší je hodnota AR indexu, tím je větší shoda mezi porovnávanými datovými celky.

2.4.2 Interní validační kritéria

Interní validační kritéria se využívají pro analýzu vnitřních charakteristik shluků.

Dendrogram

Jedním z postupů stanovení optimálního počtu shluků je heuristický přístup. Nejjednodušším příkladem je navržení počtu shluků na základě dendrogramu, v němž mohou být v některých případech znázorněny výrazné shluky.

Globální pravidla

Dalším z možných postupů je použití vybraných globálních pravidel pro stanovení počtu shluků. Pro globální stanovení jsou uvedeny indexy

$$I_1 = (\sum_B/(k-1))/(\sum_W/(n-k)), \quad (2.21)$$

kde Σ_B je meziskupinový součet čtvercových vzdáleností centroidů jednotlivých shluků od centroidu všech objektů a Σ_W je vnitroskupinový součet těchto vzdáleností jednotlivých objektů od svých centroidů (vychází se z analýzy rozptylu).

$$I_2 = (\Gamma - \Delta)/(\Gamma + \Delta), \quad (2.22)$$

kde Γ označuje počet konkordantních (shodných) srovnání a Δ počet diskordantních srovnání (je-li vnitroskupinová nepodobnost menší než meziskupinová, pak je srovnání konkordantní, je-li větší, je diskordantní).

$$I_3 = (\mathrm{D_W} - D_{min})/(\mathrm{D_{max}} - D_{min}), \quad (2.23)$$

kde D_W je součet všech vnitroskupinových nepodobností při rozdělení objektů do *k* shluků, D_{min} (resp. D_{max}) je součet minimálních (resp. maximálních) nepodobností.

Počet shluků se na základě globálních pravidel stanovuje tak, že v případě indexů I_1 a I_2 je to maximální hodnota z vypočítaných hodnot a v případě indexu I_3 minimální.

Informační kritéria

Další možností, jak stanovit optimální počet shluků je využití informačního kritéria. Rozlišují se dvě, přičemž jedno je založeno na bayesovském přístupu. To bývá

označováno zkratkou BIC (*Bayesan Information Criterion*) a nazývá se Schwarzovo bayesovské kritérium. Je vyjádřeno vztahem

$$BIC(k) = -2\sum_{h=1}^{k} \xi_h + w_k ln\,(n), \tag{2.24}$$

kde ξ_h je charakteristika *h*-tého shluku, *n* je celkový počet objektů v souboru a w_k se vypočítá podle vzorce

$$w_k = k\left(2m^{(1)} + \sum_{l=1}^{m^{(2)}} (K_l + 1)\right), \tag{2.25}$$

kde $m^{(1)}$ je počet spojitých proměnných, $m^{(2)}$ je počet kategoriálních proměnných a K_l je počet kategorií *l*-té proměnné.

Druhé informační kritérium je Akaikovo, které se označuje zkratkou AIC (*Akaike Information Criterion*). Jeho hodnota se vypočte na základě vztahu

$$AIC(k) = -2\sum_{h=1}^{k} \xi_h + 2w_k. \tag{2.26}$$

Výše zmíněná kritéria se využívají pro stanovení počtu shluků v systému SPSS, a to v proceduře dvoukrokové shlukové analýzy. [4]

Davies - Bouldin index

Davies - Bouldin (DB) index je poměrovou funkcí sumy vnitřního rozložení shluku a mezishlukové distribuce. DB index je definován jako:

$$DB = \frac{1}{N}\sum_{i,j=1}^{N} \max_{i \neq j}\left\{\frac{S_i + S_j}{d_{i,j}}\right\}, \tag{2.27}$$

kde *N* značí počet shluků, S_i a S_j jsou jednotlivé shluky a $d_{i,j}$ označuje vzdálenost mezi středy shluků. Čím je hodnota DB indexu nižší, tím je dosaženo optimálnějšího shluku. [21]

Calinski - Harabasz index

Callinski – Harabasz (CH) index je definován jako

$$CH = \frac{trace(S_B)}{trace(S_W)} \cdot \frac{n_p - 1}{n_p - k}, \tag{2.28}$$

kde S_B je mezishluková matice rozptylu, S_W matice rozptylu uvnitř shluku, n_p počet bodů ve shlucích a *k* počet shluků. Při porovnání shlukovacích algoritmů touto metodou udává maximální hodnota Calinski – Harabasz indexu nejvhodnější algoritmus. [20]

Dunn index

Dunn index je definován jako

$$Dunn = \frac{d_{min}}{d_{max}}, \tag{2.29}$$

kde d_{min} je minimální vzdálenost mezi body patřícími rozdílným shlukům a d_{max} je maximální vzdálenost mezi jakýmikoliv dvěma body stejného shluku. Nejvyšší hodnota Dunnova indexu značí optimální shlukovací algoritmus. [20]

Index siluety

Index siluety (*Silhouette index*) *S* představuje průměrnou šířku siluety všech datových bodů. Je jednou ze základních metod validace shluků. Silueta se určuje pro každý objekt, index se mění od -1 do 1, kde vyšší číslo určuje lepší míru úspěšného shlukování. [22]

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{\max(a_i, b_i)}, \tag{2.30}$$

kde a_i reprezentuje průměrnou vzdálenost bodu x_i od ostatních bodů ve stejném shluku, b_i představuje minimální průměrnou vzdálenost bodu x_i od bodů v ostatních shlucích.

GAP metoda

Tato metoda diferenční analýzy je založena na sledování změn *W(k).* Účelem je najít způsob standardizace porovnání $log(W_k)$ při rostoucím *k* s nulovým referenčním rozložením dat, tj. rozložením s nulovým zřetelným shlukováním.

Suma párových vzdáleností všech bodů ve shluku *r* je definována jako

$$D_r = \sum_{i,i' \in C_r} d_{ii'}, \tag{2.31}$$

kde $d_{ii'}$ označuje vzdálenost mezi pozorovanými body shluku *i* a *i´*, C_r označuje indexy pozorování ve shluku *r*.

$$W_k = \sum_{r=1}^{k}\frac{1}{2n_r}D_r \tag{2.32}$$

Jestliže vzdálenost *d* je kvadrát euklidovské vzdálenosti, potom W_k je suma vnitřních rozložení shluků. Odhad optimálního počtu shluků je uložen v hodnotě *k*, pro niž $log(W_k)$ je co nejvíce pod referenční křivkou. GAP statistika je popsána rovnicí

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k), \tag{2.33}$$

kde $E_n^*\{\log(W_k)\}$ označuje předpokládanou hodnotu $\log(W_k)$ referenčního nulového rozložení. Odhadované *k* bude mít hodnotu maximální $Gap_n(k)$. [25]

RMSSD index

Validační index RMSSD (*Root-Mean-Square Standard Deviation*) označuje odchylku všech proměnných ve shluku. Tento index měří v každém kroku homogenitu vytvářeného shluku. Jelikož účelem shlukování je identifikace homogenních seskupení, proto čím je RMSSD index nižší, tím je výsledek shlukování optimálnější. [24]

$$RMSSD = \sqrt{\frac{SS_w}{\sum_{\substack{i=1\dots nc \\ j=1\dots d}} (n_{ij} - 1)}}, \tag{2.34}$$

kde

$$SS_w = \sum_{\substack{i=1\dots nc \\ j=1\dots d}} \sum_{k=1}^{n_{ij}} (x_k - \overline{x_j}),^2 \tag{2.35}$$

n_c je počet shluků, d rozměr, n_{ij} označuje pořadí prvku v i-tém shluku j-tého rozměru.

RS index

RS (*R Squared*) index zjišťuje podobně jako RMSSD index homogenitu mezi shluky. Hodnoty RS indexu se pohybují v intervalu <0; 1>, kde 0 je označením pro shodné shluky, zatímco 1 indikuje významnou rozdílnost mezi zkoumanými shluky. [24]

$$RS = \frac{SS_t - SS_w}{SS_t}, \tag{2.36}$$

kde SS_w je vypočteno na základě rovnice (2.31) a SS_t je popsáno rovnicí (2.33).

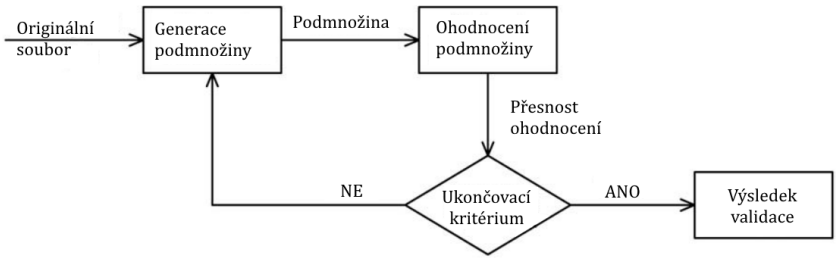
$$SS_t = \sum_{j=1}^{d} \sum_{k=1}^{n_j} \left(x_k - \overline{x_j}\right)^2 \tag{2.34}$$

2.5 Selekce příznaků

Selekce příznaků (*feature selection*) je jednou z důležitých a často využívaných technik ve strojovém učení. Redukuje počet charakteristických vlastností, odstraňuje irelevantní, redundantní nebo zašuměná data a tím výrazně zlepšuje vlastnosti analýzy: urychluje shlukování, zvyšuje přesnost a srozumitelnost výsledků. Selekce příznaků se využívá v mnoha vědních oborech, např. při rozpoznávání znaků, vyhledávání obrazů, v zákaznickém servise, při detekci vniknutí nebo genotypové analýze.

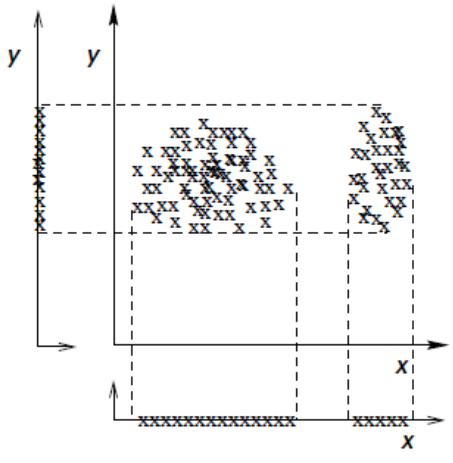
Při selekci příznaků dochází k procesu, který vybere z původních příznaků určitou podmnožinu. Její optimalita se měří pomocí hodnotícího kritéria. S rostoucím rozměrem definičního oboru roste zároveň i počet příznaků N. Nalezení optimální podmnožiny příznaků je obvykle těžko proveditelné. Typický postup při selekci příznaků se sestává ze čtyř základních kroků, které jsou zobrazeny na obrázku 2.6. Jde o

generování podmnožiny příznaků, její ohodnocení, ukončovací kritérium a kontrolu výsledku. [17], [19]

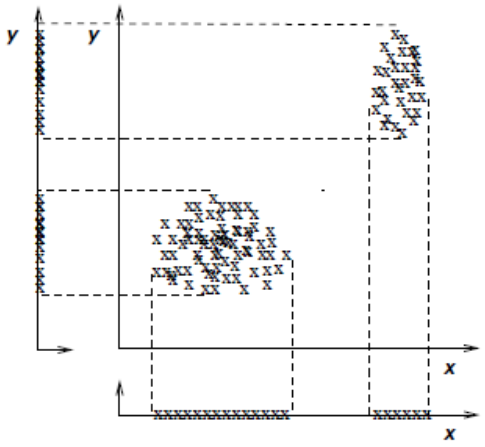


Obr. 2.6 Proces selekce příznaků [17]

Algoritmy pro selekci příznaků využívající různá hodnotící kritéria lze všeobecně rozdělit do následujících třech kategorií: filtrovací model (*filter model*), obálkový model (*wrapper model*) a hybridní model (*hybrid model*). Filtrovací model využívá při hodnocení podmnožiny příznaků základní charakteristiky dat, selekci příznaků používá jen pro předzpracování, které je nezávislé na kterémkoliv dolovacím algoritmu. Filtrovací metoda je méně časově náročná, avšak méně efektivní. Obálkový model zahrnuje jeden předem definovaný klasifikátor, jehož využívá jako hodnotícího kritéria. Vybírá podmnožinu příznaků tak, aby se zlepšil jeho výkon podle určitého kritéria. V porovnání s filtrovacími modely je obálkový model více časově náročný, ale také efektivnější. Hybridní model spojuje výhody filtrovacího a obálkového modelu využitím jejich různých hodnotících kritérií v různých fázích hledání příznaků. [17]



Obr. 2.8 Irelevantní příznaky [18]

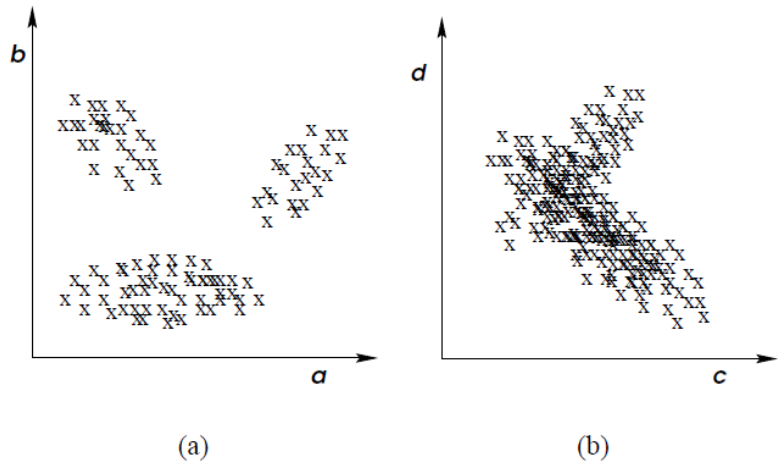


Obr. 2.7 Redundantní příznaky [18]

Tradiční algoritmy pro klasifikaci využívající selekci příznaků se díky chybějící informaci o třídě nehodí pro klasifikaci objektů učením bez učitele. Pro jejich

klasifikaci se obvykle používá snižování rozměru úlohy nebo metody pro selekci příznaků, např.: analýza hlavních komponent (*Principal Components Analysis, PCA*), Karhunen-Loéveho transformace nebo singulární rozklad (S*ingular Value Decomposition, SVD*).

Obrázek 2.8 ukazuje příklad nadbytečnosti příznaků při učení bez učitele. Data mohou být seskupena do stejných shluků podle příznaku *x* nebo podle příznaku *y*. Poté se jeden z příznaků stává nadbytečným. Obrázek 2.7 ukazuje příklad, kdy jsou příznaky irelevantní. Příznak *y* nemá podíl na rozlišení jednotlivých shluků. Při jeho použití je možné určit pouze jeden shluk. Irelevantní příznaky mohou vést k chybným výsledkům shlukování. Reálná situace však může být komplexnější, než jak je naznačeno na obrázcích 2.7 a 2.8. Tato situace je názorně ukázána na obrázku 2.9, kde výsledné shluky byly určeny použitím dvou podmnožin příznaků {a, b} a {c, d}. Odlišné podmnožiny příznaků vedou k vytvoření odlišných shluků. [18]



Obr. 2.9 Shluky z různých podmnožin příznaků [18]

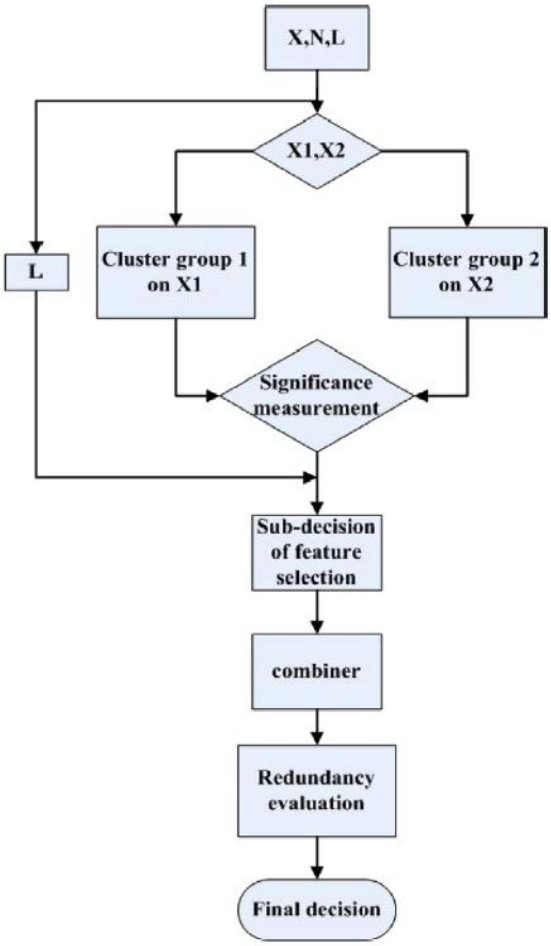
Jednou z metod, která dokáže překonat zmíněné nedostatky je metoda hybridního výběru příznaků pro učení bez učitele (*Hybrid Feature Selection Scheme for Unsupervised Learning, HFS*) která je podrobněji popsána v kapitole 2.5.1 a v kapitole 4.4 je experimentálně ověřen její přínos pro úspěšnost shlukové analýzy. Dále jsou v této kapitole stručně popsány další metody selekce příznaků v úlohách bez učitele, které byly vybrány z velkého množství existujících metod. Jsou zde prezentovány tři algoritmy typu *Filter* (LVF, Focus, Relief), dva typu *Wrapper* (B&B, SFFS) a šest algoritmů dalších typů.

2.5.1 HFS

Metoda hybridního výběru příznaků pro učení bez učitele HFS generuje pro následné shlukování dva náhodně zvolené podprostory, kombinuje různé typy shlukové

analýzy k získání konečné množiny dílčích označení shluků. Ty se využijí k selekci příznaků na základě významných měření. Pro zlepšení kvality vybraných příznaků odstraňuje na základě relevance příznaků ty z nich, které jsou nadbytečné.

Metoda HFS zahrnuje dva aspekty: (1) ohodnocení významu příznaku - bere v úvahu přínos každého příznaku pro výsledek shlukování; (2) ohodnocení nadbytečnosti - na základě podobnosti příznaků zachovává nejvýznamnější a nezávislé příznaky. Vývojový diagram HFS algoritmu je ukázán na obrázku 2.10 a níže je popsán pseudokódem samotný algoritmus.



Obr. 2.10 Vývojový diagram algoritmu HFS [19]

Pseudokód algoritmu HFS

Vstup: prostor příznaků *X*, počet shluků *N*, maximum iterací *L*

Výstup: rozhodnutí o výběru příznaků

(1) Opakuj tak dlouho, dokud nezískáš konečnou množinu dílčích výsledků

for k = 1 : L do:

(1.1) Náhodně rozděl původní prostor s příznaky do dvou podprostorů X_1, X_2

(1.2) a) Na data v podprostoru X_1 použij první shlukovací algoritmus, výsledek (označení shluků) ulož do vektoru *Rk1*

b) Na data v podprostoru X_2 použij druhý shlukovací algoritmus, výsledek (označení shluků) ulož do vektoru *Rk2*

(1.3) a) Pomocí jednoho z algoritmů (LCC, SU, DB) ohodnoť význam každého příznaku *F* proti *Rk1* a označ ho *RF(k1)*

b) Pomocí jednoho z algoritmů (LCC, SU, DB) ohodnoť význam každého příznaku *F* proti *Rk2* a označ ho *RF(k2)*

// Sluč všechny množiny řazení *RF* do jedné

(2) $RF^{prefinal} = combiner\{ RF_{(1)}, RF_{(2)}, \dots, RF_{(2L)}\}$

(3) Ohodnoť nadbytečnost získaných příznaků

(4) Vrať konečnou množinu příznaků RF^{final}

V kroku (1.3) výše popsaného pseudokódu je ohodnocen význam každého příznaku s ohledem na jeho přínos pro výsledek shlukování. Pro měření korelace mezi příznaky a výsledkem shlukování se využívá několik přístupů: lineární korelace a přístup založený na výpočtu entropie. Pro výpočet ohodnocení relevance příznaků (tzv. *rankování*) bylo využito lineárního korelačního koeficientu LCC (*Linear Correlation Coefficient*) a metody souměrné nejistoty SU (*Symmetrical Uncertainity*). Metoda SU patří mezi jedny z nejvíce efektivních přístupů pro selekci příznaků založených na výpočtu entropie.

a) **Lineární korelační koeficient**

Metoda LCC pokládá obě hodnoty příznaků a štítků (názvů tříd) jednotlivých instancí za proměnné a studuje korelace mezi těmito proměnnými. Metoda je definována vzorcem (2.35). F_k označuje *k*-tý příznak a R_k *k*-tou identifikaci shluku vzniklou jako výsledek shlukování.

$$LCC(F_k, R_k) = \frac{cov(F_k, R_k)}{\sigma(F_k)\sigma(R_k)}, \tag{2.35}$$

kde $\sigma(R_k)$ znamená standardní odchylku *k*-té identifikace shluků a $cov(F_k, R_k)$ je kovariance mezi F_k a R_k, která je dána vzorcem (2.36):

$$cov(F_k, R_k) = \frac{\sum_{i=1}^{N}(\widehat{F_k} - d_{ik})(\widehat{R_k} - R_i)}{N}, \tag{2.36}$$

kde R_i znamená označení přiřazení třídy instanci d_i a $\widehat{R_k}$ je průměrná hodnota označení tříd všech instancí. Standardní odchylka $\sigma(F_k)$ je definována jako:

$$\sigma(F_k) = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{F_k} - d_{ik})^2}{N}}, \tag{2.37}$$

kde $\widehat{F_k}$ je průměrná hodnota příznaku F_k a je dána rovnicí (2.38):

$$\widehat{F_k} = \frac{\sum_{i=1}^{N} d_{ik}}{N}. \tag{2.38}$$

b) Metoda souměrné nejistoty

Metoda SU je definována jako:

$$SU(F_k, R_k) = 2\left[\frac{IG(F_k|R_k)}{H(F_k) + H(R_k)}\right], \tag{2.39}$$

kde $H(R_k)$ je entropie tříd instancí a $IG(F_k|R_k)$ se nazývá informační zesílení, jež je možné vypočítat pomocí vzorce (2.40):

$$IG(F_k|R_k) = H(F_k) - H(F_k|R_k), \tag{2.40}$$

kde $H(F_k)$ je entropie příznaku F_k a $H(F_k|R_k)$ je podmíněná entropie příznaku F_k. Jestliže $\Omega(F_k)$ představuje všechny možné hodnoty příznaku F_k a $\Omega(R_k)$ představuje všechny hodnoty označení tříd, pak $H(F_k)$ a $H(F_k|R_k)$ lze vyjádřit vzorci:

$$H(F_k) = -\sum_{F'_k \in \Omega(F_k)} P(F'_k)\log\big(P(F'_k)\big), \tag{2.41}$$

$$H(F_k|R_k) = -\sum_{R'_k \in \Omega(R_k)} P(R'_k)\left[\sum_{F'_k \in \Omega(F_k)} P(F'_k|R')\log\big(P(F'_k|R')\big)\right], \tag{2.41}$$

kde $P(F'_k)$ označuje pravděpodobnost, že hodnota příznaku F_r bude shodná s hodnotou příznaku F'_r. $P(F'_k|R)$ udává pravděpodobnost, že hodnota příznaku F_r bude shodná s hodnotou příznaku F'_r za podmínky, že instance jsou přiřazeny do třídy R'.

$$P(F'_k) = \frac{\sum_{i=1}^{N} \delta(d_i, F'_k)}{N}, \tag{2.42}$$

kde

$$\delta(d_i, F'_k) = \begin{cases} 1, & pokud\ d_i = F'_k \\ 0, & v\ ostatních\ případech \end{cases}. \tag{2.43}$$

V dalším kroku (2) pseudokódu jsou zkombinována všechna řazení *RF* do jedné množiny „ranků“ $RF^{pre\text{-}final}$. Vypočte se průměrná relevance příznaků (*rank*) *AR* pro každý příznak:

$$\mathrm{AR(j)} = \frac{\sum_{\mathrm{k=1}}^{\mathrm{M}} \mathrm{rank_{(k)}(j)}}{\mathrm{M}}, \tag{2.44}$$

kde *M* znamená počet instancí (řádků) a $rank_{(k)}(j)$ představuje relevanci příznaku *j* v ohodnocení příznaku $RF_{(k)}$ vypočteném v kroku (1.3). Výsledná hodnota *AR* se odečte od hodnoty celkového počtu příznaků *n* a tím se získá kumulativní *rank CR*, podle kterého se následně ohodnotí všechny příznaky:

$$CR(j) = n - AR(j). \tag{2.45}$$

Nejvíce důležitý příznak má ohodnocení (*rank*) rovno jedné. To znamená, že příznaky jsou ohodnoceny sestupně podle své významnosti.

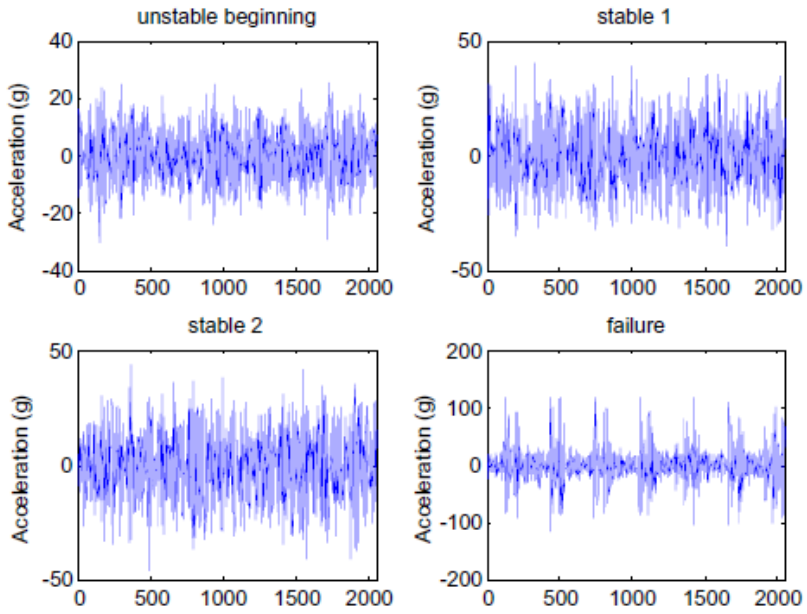
V dalším kroku (3) dochází k eliminaci redundantních veličin. Pro první vybraný příznak je spočítána korelace vůči ostatním příznakům. Pokud pro nějaký pár vyjde hodnota korelace větší, než zadaná prahová hodnota, je daná veličina (příznak) odstraněna.

V kroku (4) je po odstranění redundantních veličin získána konečná množina příznaků RF^{final}.

Metoda selekce příznaků HFS [19] vychází z metody selekce příznaků FRMV popsané v článku [9], odkud jsou čerpány výše použité vzorce.

Experimentální ověření metody HFS autory článku

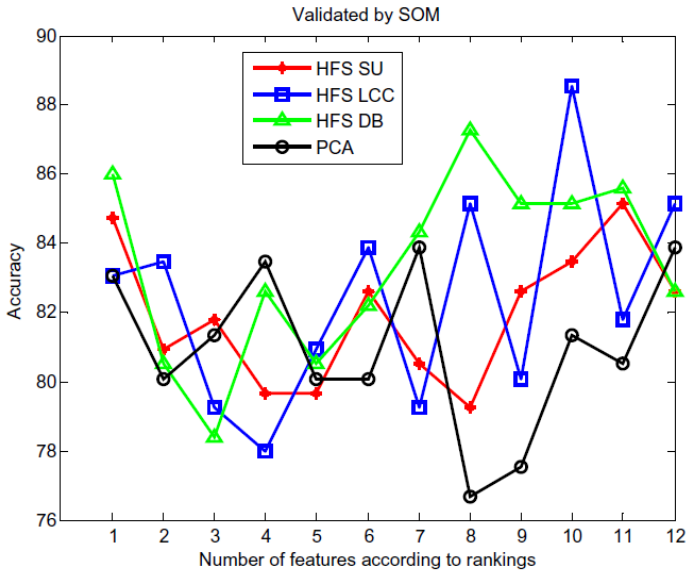
Metoda hybridního výběru příznaků pro učení bez učitele HFS byla porovnána při diagnostice poruch valivých ložisek s ostatními metodami pro výběr příznaků:



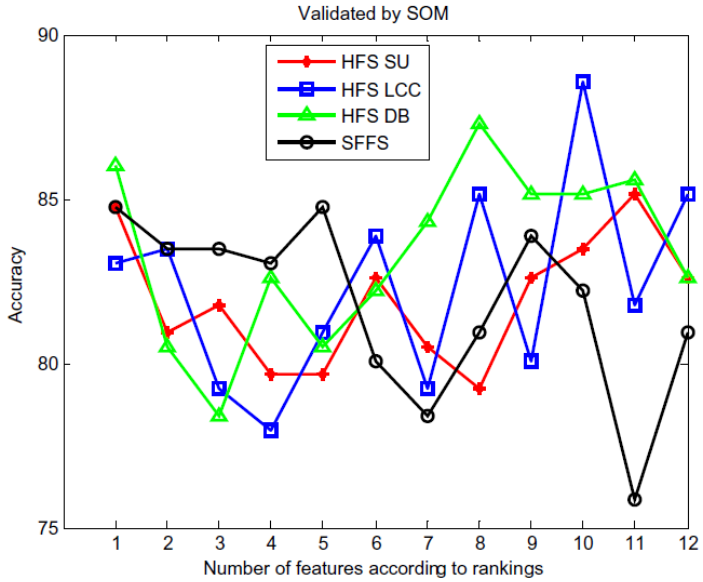
Obr. 2.11 Působení vibračního signálu na kuličkové ložisko [19]

HFS SU (HFS s hodnotícím kritériem SU (*Symmetrical Uncertainity*)), HFS LCC (HFS s lineárním korelačním kritériem LCC (*Linear Correlation Coefficient*)), HFS DB (HFS s DB indexem), SFFS (*Forward Search Feature Selection*) a analýzou hlavních komponent PCA (*Principal Component Analysis*). K porovnání výkonu metod pro klasifikaci na základě výběru příznaků byly využity samoorganizující se mapy SOM.

K testování byla použita sada jednořadých kuličkových valivých ložisek. Při různé frekvenci otáčení této sady bylo zaznamenáváno chování ložisek. Obrázek 2.11 ukazuje čtyři fáze testu: (1) nestabilní počáteční stav; (2) první stabilní segment; (3) druhý stabilní segment a (4) defekt vnitřního kroužku ložiska.



Obr. 2.12 Srovnání přesnosti klasifikace metody HFS a SFFS [19]



Obr. 2.13 Srovnání přesnosti klasifikace metody HFS a PCA [19]

Obrázky 2.12 a 2.13 ukazují výsledky výše zmíněného testu. Při srovnání přesnosti klasifikace metod pro výběr příznaků založených na metodě HFS vychází, že metoda HFS LCC dosáhla nejvyšší přesnosti 88,5% s prvními deseti příznaky, zatímco metody HFS DB a HFS SU dosáhly své největší přesnosti 87,29% a 85,17% s prvními osmi, respektive jedenácti příznaky.

Na obrázku 2.12 je srovnání přesnosti skupin metod založených na HFS s metodou SFFS. Metoda SFFS dosahuje přesnosti 84,75%.

Obrázek 2.13 srovnává metody HFS s metodami založených na PCA (*Principal Component Analysis*). Vyplývá z něj, že metody HFS LCC a HFS DB dosáhly v porovnání s PCA metodami (přesnost 86,02%) lepší přesnosti bez ohledu na počet zvolených příznaků.

Skupina metod pro výběr příznaků založených na metodě HFS prokázala, že dosahuje lepší přesnosti klasifikace s méně příznaky. HFS je robustní a efektivní metodou pro výběr příznaků, výrazně zlepšuje výkon klasifikace s učením bez učitele. Její výkonnost však závisí na výběru hodnotícího kritéria. [19]

2.5.2 LVF

LVF (*Las Vegas Filter*) algoritmus opakovaně generuje náhodné podmnožiny *S* příznaků z celkové množiny příznaků. Jestliže *S* obsahuje méně příznaků, než aktuálně vybraná nejlepší podmnožina, poměr nekonzistentnosti redukovaných dat *S* je porovnán s poměrem nekonzistentnosti aktuální nejlepší podmnožiny. Jestliže je *S* alespoň tak konzistentní jako aktuální nejlepší podmnožina, nahradí ji. LVF je tedy ovlivňována konzistentností a umí pracovat se zašuměnými daty, pokud je a priori znám přibližný stupeň jejich zašumění. [12]

2.5.3 RELIEF

Relief je jednoduchý algoritmus, který určuje relevantní příznaky pomocí statistických metod. Pracuje s náhodně vybranými instancemi (vzorky) z trénovacích dat. Pro každou instanci zjišťuje nejbližší instanci ve stejné třídě (*nearest hit*) a jiné třídě (*nearest miss*). Výsledkem je původní datová struktura, avšak s přiřazením vah jednotlivým příznakům. [27]

2.5.4 FOCUS

Algoritmus důkladně prozkoumává prostor podmnožiny příznaků, dokud nenalezne minimální kombinaci příznaků, jež rozdělí trénovací data do tříd. To se stane tehdy, když každá kombinace hodnoty příznaku je přiřazena jedné jedinečné třídě. Tento stav se nazývá „*min-features bias*“. Po selekci příznaků je z konečné množiny příznaků sestaven rozhodovací strom. [12]

2.5.5 SFG/SBG

SFG (*Sequential Forward Generation*) algoritmus iterativně přidává příznaky do počáteční podmnožiny příznaků, čímž zlepšuje ohodnocení příznaků v počáteční množině příznaků. SBG (*Sequental Backward Generation*) algoritmus pracuje přesně na opačném principu než SFG algoritmus. [27]

2.5.6 SFFS

Algoritmus SFFS (*Sequential Floating Forward Selection*) je založen na algoritmu sekvenčního dopředného vyhledávání SFS (*Sequential Feature Selection*), který navíc využívá zpětné vyhledávání, tzv. „*backtracking*“. SFS začíná s prázdnou množinou a postupně přidává vhodné příznaky. Při zpětném vyhledávání odebírá nevyhovující příznaky. *Backtracking* se opakuje tak dlouho, dokud algoritmus nenalezne množinu s požadovaným počtem příznaků. [29]

2.5.7 B&B

Metoda větvení a mezí (*Branch & Bound*) je optimální vyhledávací algoritmus. Principem je systematické procházení všech potencionálních řešení, velké podmnožiny nevhodných kandidátů se vyřazují najednou. Vyhledávání se zastaví v každém vyhodnocovacím uzlu, pokud má daný uzel hodnotu menší, než uživatelem nastavený práh β. Tím dojde k „prořezání“ původních větví. [27]

2.5.8 MCFS

MCFC (*Multi-Cluster Feature Selection*) je metoda selekce příznaků v úlohách bez učitele, která se využívá ve spektrální analýze k uchování co nejoptimálnější víceshlukové datové struktury. Metoda MCFS poskytuje řešení optimalizačního problému, který se týká L1-regulace nejmenších čtverců a zřídka se vyskytujících problémů. [28]

2.5.9 FSFS

FSFS (*Feature Similarity Feature Selection*) zjišťuje podobnost příznaků metodou maximálního zhuštění informací. Pracuje ve dvou krocích. Nejdříve rozdělí původní množinu příznaků na určitý počet homogenních podmnožin (shluků) a dále provádí výběr reprezentativních příznaků z každého shluku. Proces se opakuje až do doby, kdy všechny příznaky shluku jsou buď vybrány, nebo odstraněny. [30]

2.5.10 CPFS

CPFS (*Convex Principal Feature Selection*) je metoda založena na metodě analýzy hlavních komponent PCA. Efektivně vybírá příznaky korespondující s prvními *k* hlavními komponentami. V porovnání s ostatními PCA metodami dosahuje nízké rekonstrukční chyby datové matice a zároveň minimalizuje redundantní příznaky. [31]

2.5.11 UDFS

Dosud existující algoritmy vybírají příznaky při zachování datové struktury celé množiny příznaků. UDFS (*Unsupervised Discriminative Feature Selection*) je nový

algoritmus, který na rozdíl od stávajících algoritmů vybírá z datové množiny charakteristické příznaky pro učení bez učitele. Metoda při výběru příznaků spojuje diskriminativní analýzu s $l_{2,1}$-*norm* minimalizací. UDFS je *data-driven* metoda, tj. není vhodná pro výběr příznaků z vícerozměrných dat a velmi malých souborů. [32]

2.6 Požadavky na metody shlukové analýzy

- Škálovatelnost na velké datové soubory
- Schopnost zacházení s různými typy atributů
- Nalezení shluků libovolného tvaru
- Požadavek alespoň minimální znalosti oboru pro určení vstupních parametrů
- Schopnost vypořádat se s šumem v datech
- Schopnost vypořádat se s odlehlými objekty
- Nezávislost na pořadí vstupních záznamů
- Časová optimalizace
- Schopnost práce s vícerozměrnými daty
- Interpretovatelnost a použitelnost výsledků [10]

3 METODY SHLUKOVÉ ANALÝZY

V literatuře je uváděno mnoho typů metod shlukové analýzy a na jejich klasifikaci existuje mnoho pohledů. Je obtížné provést jejich přesné zařazení do kategorií, protože jednotlivé kategorie se mohou navzájem překrývat a daná metoda může mít charakteristické vlastnosti několika dalších shlukovacích metod. V poslední době byla navržena řada nových algoritmů. Některé modifikují tradiční metody shlukové analýzy (metody rozkladu a hierarchické metody), jiné se vydávají novými směry. Vzniká tak terminologický problém, co vše by se ještě mělo nazývat shlukovou analýzou a co už by se mělo označovat jinak. Protože nové metody jsou vyvíjeny často mimo sféru statistické analýzy dat (jde např. o vědní obory v oblasti informatiky, jako je vyhledávání informací v rozsáhlých databázích či rozpoznávání vzorů), převažuje v literatuře termín shlukovací techniky (*clustering techniques*). Avšak někteří autoři používají pojem shluková analýza pro širší okruh metod, než jsou tradiční. [1], [4]

Metody shlukové analýzy můžeme rozčlenit do následujících hlavních kategorií:

Metody rozkladu (*Partitioning methods*) - někdy též zvané nehierarchické metody (*flat*) vytvářejí konkrétní počet shluků. Přiřazení ke shlukům je buď jednoznačné, nebo se počítá míra příslušnosti jednotlivých objektů ke shlukům. Příkladem metod rozkladu je metoda *k*-průměrů a její modifikace (*k*-medoidů, *k*-modů, *k*-histogramů). Míru příslušnosti ke shlukům je možné zjistit pomocí fuzzy shlukové analýzy. Více v kapitole 3.1. [4]

Hierarchické metody (*Hierarchical methods*) - vytváří hierarchickou posloupnost rozkladů předložených dat neboli dendogram. Tyto metody se dále dělí na *aglomerativni* a *divizní*. Aglomerativní algoritmy postupně spojují objekty a jejich shluky až do vzniku jednoho shluku. Divizní algoritmy naopak postupně rozdělují množinu (shluk) objektů na podmnožiny. Více v kapitole 3.2. [5]

Metody založené na hustotě (*Density-based methods*) – jejich hlavní podstatou je neustálý růst daného shluku až do té doby, než hustota bodů (počet objektů) v sousedství překročí určitou mezní hodnotu. Těchto metod se využívá k filtrování šumu a ke zjišťování odlehlých objektů a shluků libovolných tvarů. Příkladem metod založených na hustotě jsou DBSCAN, OPTICS, DBCLASD a DENCLUE. Více v kapitole 3.3. [1]

Metody založené na mřížce (*Grid-based methods*) – rozdělují datový prostor na konečný počet buněk tvořících mřížkovou strukturu. Na této mřížkové struktuře jsou prováděny veškeré operace. Hlavní výhodou tohoto přístupu je rychlé zpracování, které

je obvykle závislé na počtu buněk v každé dimenzi rozděleného prostoru. Příkladem metod založených na mřížce jsou algoritmy STING, CLIQUE nebo WaveCluster. Více v kapitole 3.4. [11]

Metody založené na modelu (*Model-based methods*) – předpokládají pro každý shluk model a k němu hledají nejlepší přiřazení dat. Na základě funkce hustoty, která odráží prostorové rozmístění bodů, umožňují lokalizovat shluky. Na základě statistik také automaticky určí počet shluků. Příkladem metod založených na modelu jsou rozhodovací stromy (algoritmus COBWEB) a algoritmus SOON založený na samo-organizující se neuronové síti SOM. Více v kapitole 3.5. [11]

Výše popsané metody jsou stručně shrnuty v tabulce 3.1:

Metoda	Charakteristiky metod
Metody rozkladu	• Umí nalézt vzájemně uzavřené shluky pro sférický tvar • Založeny na výpočtu vzdálenosti • K reprezentaci středu shluku využívají průměru, medoidu, … • Efektivní pro malé až středně veliké datové sety
Hierarchické metody	• Shluky je možné hierarchicky rozložit • Neumí opravit chybné seřazení nebo rozklad • Mohou zahrnovat i jiné shlukovací techniky, jako mikroshlukování nebo vazební (linkové) metody
Metody založené na hustotě	• Umí nalézt nepravidelné shluky • Umožňují odfiltrovat odlehlé objekty • Definují shluky jako regiony s velkou hustotou objektů • Vyžadují pro každý bod minimální počet bodů v jeho sousedství
Metody založené na mřížce	• Používají mnohovrstevnou mřížkovou strukturu • Rychlý čas zpracování nezávisí na počtu objektů, ale na velikosti mřížky
Metody založené na modelu	• Optimalizují vazbu mezi daty a matematickým modelem • Umí najít charakteristické rysy každého objektu

Tab. 3.1 Charakteristiky shlukovacích metod [1]

3.1 Metody rozkladu

Nejjednodušší a nejzákladnější variantou shlukové analýzy jsou metody rozkladu, které optimalizují rozklad množiny objektů do předem zadaného počtu skupin nebo shluků. Právě znalost počtu shluků je výchozí znalostí metod rozkladu.

Mějme skupinu dat *D* s počtem *n* objektů a *k* počtem shluků. Algoritmus rozkladu na počátku rozřadí objekty do *k* částí ($k \leq n$), kde každá část reprezentuje shluk. Shluky jsou vytvářeny k optimalizaci kritérií rozkladu, takže objekty uvnitř shluku se vzájemně podobají a liší se od objektů v ostatních shlucích.

Jelikož je výpočet všech možných skupin shluků velmi náročný úkon i pro počítačové zpracování, určité heuristické metody výpočtu využívají pro zjednodušení iterativní optimalizaci. Používají tedy různá relokační schémata, která iterativně přerozdělují objekty mezi jednotlivými *k* shluky. Na rozdíl od tradičních hierarchických metod, v nichž shluky nejsou po svém vzniku již dále přerozdělovány, relokační algoritmy se shluky dále pracují a zdokonalují je.

Nejznámější a běžně používané metody rozkladu jsou metody *k*-průměrů, *k*-medoidů a jejich variace.

Jedním z dalších přístupů pro rozklad skupiny dat je koncepční přístup, který porovnává shluky s určitým modelem, jehož pravděpodobnostní rozložení a parametry musí být teprve nalezeny. Jsou to tzv. pravděpodobnostní modely, které předpokládají, že data se skládají z několika druhů objektů, jejichž nejlepší umístění chceme najít. Odpovídající algoritmy jsou popsány v kapitole 3.1.1 Pravděpodobnostní shlukování. Jednoznačnou výhodou pravděpodobnostních metod je interpretovatelnost výsledných shluků. [1], [10]

3.1.1 Pravděpodobnostní shlukování

Data jsou v případě pravděpodobnostního shlukování považována za vzorky získané nezávisle z modelu, který obsahuje různé typy objektů (*mixture model*) s různým pravděpodobnostním rozložením. Hlavním předpokladem je, že datové objekty jsou generovány jednak náhodným výběrem z modelu *j* s pravděpodobností τ_j, $j = 1:k$, a dále výběrem vhodného objektu *x* ze zvoleného smíšeného modelu. Přirozeným způsobem se tak v určité oblasti vytváří shluk, který má svůj střed. Každý datový objekt shluku má kromě svých zjevných atributů také skrytý atribut „*cluster ID*“ neboli identifikátor shluku.

V průběhu let se vyvinuly různé algoritmy pravděpodobnostního shlukovaní:

EM metoda (*Expectation-Maximization method*) – dvoukroková iterativní optimalizační metoda.

SNOB – využívá smíšených modelů ve spojení s kritériem MML (*Minimum Message Length Criterion*).

AutoClass – využívá smíšený model a umožňuje velkou variabilitu pravděpodobnostního rozložení (Bernoulliovo, Gaussovo, Poissonovo rozložení, rozložení podle přirozeného logaritmu, ...).

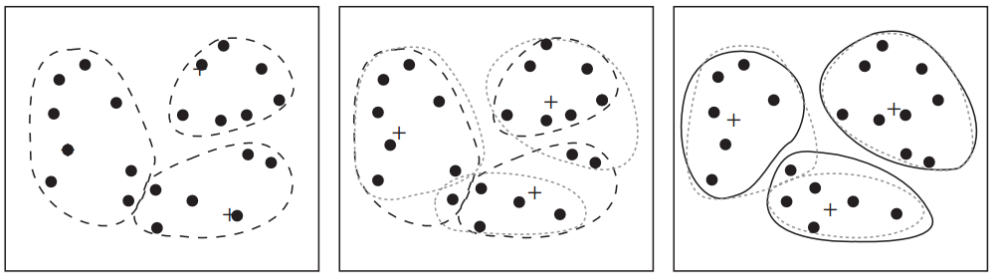
MCLUST – softwareový balíček pro hierarchické, smíšené modely shlukování a diskriminační analýzu využívající BIC (*Bayesian Information Criterion*) kritéria pro zhodnocení správnosti výběru. [10]

3.1.2 Metoda *k*-průměrů

Metoda *k*-průměrů (*k-means method*) je zdaleka nejpopulárnější shlukovací metodou *k*-shlukování používanou ve vědě a průmyslových aplikacích. Někdy se také tyto algoritmy nazývají algoritmy *k*-centroidů. Každý z k shluků C_j je totiž reprezentován vektorem průměrných hodnot jednotlivých proměnných, takzvaným centroidem. Vzdálenost mezi objektem $p \in C_j$ a centroidem c_i je zjištěna měřením euklidovské vzdálenosti mezi objektem p a centroidem c_i, definované jako minimalizace funkce

$$f_{KP} = \sum_{i=1}^{k} \sum_{p \in C_i} |p - c_i|^2, \tag{3.1}$$

kde f_{KP} je součet kvadrátů odchylek všech objektů v datovém souboru. To znamená, že pro každý objekt v každém shluku se vypočítá mocnina vzdálenosti mezi objektem a shlukem a všechny tyto vzdálenosti se sečtou.



Obr. 3.1 Metoda *k*-průměrů [1]

Jak pracuje algoritmus *k*-průměrů, je vysvětleno na příkladu, který je zdokumentován na obrázku 3.1. Je dána počáteční množina bodů ve 2D prostoru, kterou chceme rozdělit do tří shluků. Nejprve jsou zvoleny libovolné tři objekty jako středy budoucích shluků a označeny znaménkem „+“. Každý objekt je poté přiřazen do shluku podle vzdálenosti ke zvolenému středu, viz obrázek 3.1(a).

V dalším kroku jsou aktualizovány středy shluků a přerozděleny objekty do shluků. Střední hodnota každého shluku se přepočte na základě aktuálních objektů v daném shluku. Provede se výpočet vzdáleností objektů k nově vypočteným středům shluků a na jejich základě jsou objekty přiřazeny do nejbližšího shluku. Obrázek 3.1(b) zobrazuje tečkovaně nově vyznačené shluky.

Obrázek 3.1(c) zobrazuje výslednou podobu shluků, kterých bylo dosaženo iterací výše zmíněného procesu. Tento iterativní proces se nazývá iterativní relokace. [1], [10]

Jedním z důvodů velké oblíbenosti algoritmu *k*-průměrů je jeho lineární komplexnost. Dokonce, i když je počet požadavků značně velký (což je běžné), je algoritmus dobře použitelný. To je jedna z výhod oproti ostatním shlukovacím metodám (např. hierarchickým metodám), které mají nelineární závislost. Dalším důvodem je jeho snadná interpretace, jednoduchá implementace, rychlost konvergence a přizpůsobivost rozptýleným datům. Mezi zápory algoritmu patří jeho citlivost na šum a odlehlé objekty (jediný odlehlý objekt může dramaticky zvětšit kvadrát odchylky), je použitelný pouze pokud je definovaný střed a vyžaduje předem zadaný počet shluků, což není snadné, pokud nemáme o datech žádné znalosti. [12]

Existuje několik variant a modifikací metod *k*-průměrů. Liší se např. volbou inicializace algoritmu *k*-průměrů, výpočtem nepodobnosti a strategií výpočtu středu shluku. [10]

3.1.3 Metoda *k*-medoidů

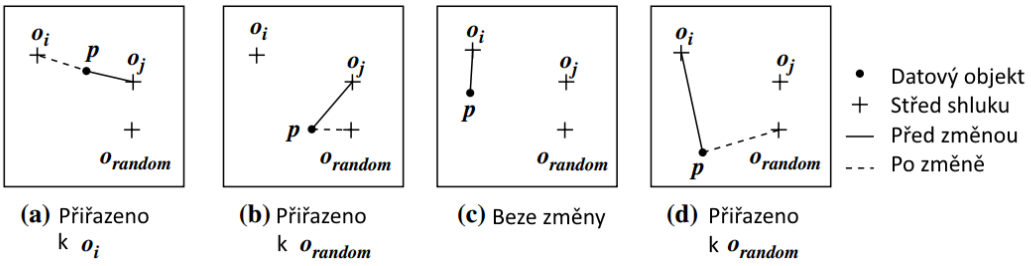
Známou realizací metody *k*-medoidů (*k-medoids method*) je algoritmus PAM (*Partitioning Around Medoids*). Stejně jako v případě metod *k*-průměrů je tento algoritmus určen pro kvantitativní proměnné a vychází z počátečního rozdělení objektů do *k* shluků. Pro každý vytvořený shluk je zjištěn medoid, což je konkrétní objekt, pro nějž je součet vzdáleností k ostatním objektům shluku minimální. To znamená, že se minimalizuje funkce

$$f_{KM} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{ih} \|\boldsymbol{x}_i - \boldsymbol{m}_h\|, \tag{3.2}$$

kde prvky $u_{ih} \epsilon \{0,1\}$ označují, zda *i*-tý objekt je či není přiřazen k *h*-tému shluku, a m_h je medoid *h*-tého shluku. Neboli, místo výběru středového objektu ve shluku jako referenčního bodu je možné vybrat jeden běžný objekt shluku jako jeho reprezentanta. K tomuto reprezentantovi se přidávají shluky, které jsou mu nejvíce podobné.

Počáteční výběr objektů je náhodný. Iterativní proces nahrazování objektů reprezentantů jinými objekty pokračuje tak dlouho, dokud se zlepšuje kvalita výsledných shluků. Kvalita je určena hodnotou funkce, která měří průměrnou nepodobnost mezi objektem a reprezentantem shluku. Při rozhodování, zda objekt o_{random} je dobrou náhradou současného reprezentanta o_j, jsou zkoumány následující čtyři možnosti, které ilustruje Obrázek 3.2.

a) p patří reprezentantovi o_j. Jestliže nahradíme objekt o_j reprezentantem o_{random} a p je nejblíže k jednomu z dalších reprezentantů o_i, kde $j \neq i$, pak je p přiřazeno k objektu o_i
b) p patří reprezentantovi o_j. Jestliže nahradíme objekt o_j reprezentantem o_{random} a p je nejblíže k objektu o_{random}, pak je p přiřazeno k objektu o_{random}
c) p patří reprezentantovi o_i, kde $j \neq i$. Jestliže nahradíme objekt o_j reprezentantem o_{random} a p je stále nejblíže k objektu o_i, pak zůstává přiřazení objektů neměnné
d) p patří reprezentantovi o_i kde $j \neq i$. Jestliže nahradíme objekt o_j reprezentantem o_{random} a p je nejblíže k objektu o_{random}, pak p je přiřazeno k objektu o_{random}



Obr. 3.2 Metoda *k*-medoidů [11]

Metoda *k*-medoidů je robustnější než metoda *k*-průměrů na přítomnost šumu a odlehlých objektů, jelikož medoid je méně ovlivnitelný odlehlými objekty a ostatními extrémními hodnotami. Metoda je ale více náročnější na počítačové zpracování. [4], [5], [11], [12]

3.1.4 Metoda *k*-modů a *k*-histogramů

Další z modifikací metody *k*-průměrů jsou algoritmy *k*-modů a *k*-histogramů, určené pro shlukování objektů charakterizovaných nominálními proměnnými. V případě algoritmu *k*-modů se používá koeficient prosté shody, viz vzorec (2.8), resp. míra nepodobnosti z něho odvozená. Tehdy je m-rozměrný vektor modálních (nejčastěji zastoupených) kategorií speciálním typem centroidu, obdobně jako vektor průměrů či mediánů.

Kromě výše uvedených postupů zaměřených na určitý typ proměnných byl navržen postup pro shlukování objektů charakterizovaných proměnnými různých typů. Nazývá se algoritmus *k*-prototypů. Shluk je reprezentován *p*-rozměrným vektorem (prototypem) vhodných charakteristik jednotlivých kvantitativních proměnných a modálními kategoriemi pro proměnné nominální. Podobnost objektů je posuzována pomocí speciálních měr nepodobnosti. [4]

3.2 Hierarchické metody

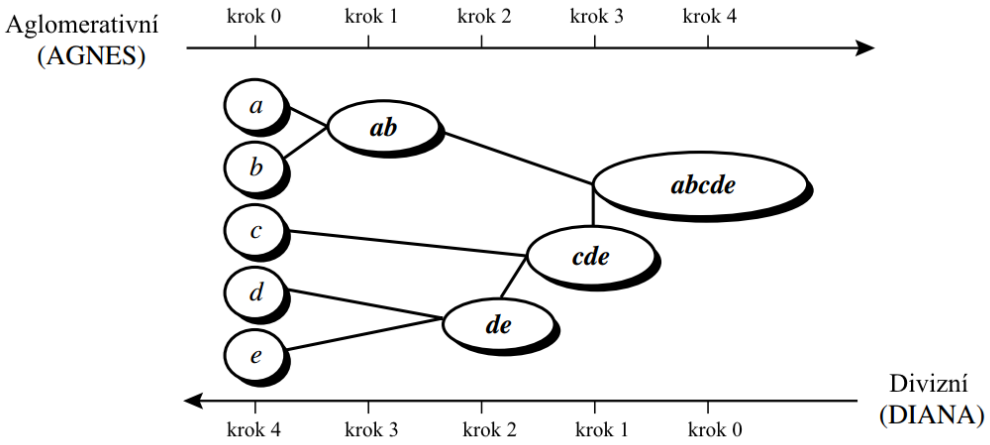
Hierarchické metody shlukování pracují na principu seskupování objektů do hierarchických shluků, tzv. dendogramů. Dendogram je běžně používán pro reprezentaci procesu hierarchického shlukování. Krok po kroku zobrazuje, jak jsou data seskupována. Hierarchické metody se dále dělí na aglomerativní (analýza podobnosti) a divizní (analýza nepodobnosti). Aglomerativní a divizní shlukování se s výhodou aplikuje na shlukování dokumentů. Divizní algoritmy se navíc využívají například v jazykovědě nebo při dolování znalostí. [3], [11]

Aglomerativní hierarchické shlukování

Tato „*bottom-up*“ (zespoda nahoru) strategie shlukování vychází ze stavu, v němž je každý objekt samostatným shlukem. Postupně se tyto shluky na základě podobnosti spojují ve větší a větší shluky, dokud nejsou všechny objekty sloučeny do jednoho shluku. Do této kategorie spadá mnoho shlukovacích hierarchických metod, které se navzájem liší pouze v definici mezishlukové podobnosti (*intercluster similarity*).

Divizní hierarchické shlukování

Tato „*top-down*“ (shora dolů) strategie shlukování pracuje na opačném principu než aglomerativní přístup shlukování. Na začátku tvoří všechny objekty jeden shluk. Algoritmus rozděluje shluk na menší a menší části, dokud každý objekt není samostatným shlukem nebo není splněna nějaká cílová podmínka (např.: je dosaženo požadovaného počtu shluků, poloměr shluku přesáhl určitou mez).



Obr. 3.3 Aglomerativní a divizní hierarchické shlukování [11]

Na obrázku 3.3 je znázorněno použití aglomerativního (*AGNES – Aglomerative NEsting*) a divizního (*DIANA – Divisive ANAlysis*) přístupu shlukování na množině dat

s pěti objekty $\{a, b, c, d, e\}$. Na začátku umístí AGNES každý objekt do samostatného shluku. Shluky jsou poté krok po kroku podle zadaného kritéria (např. minimum euklidovské vzdálenosti mezi dvěma objekty z různých shluků) slučovány. Tento proces pokračuje tak dlouho, dokud nejsou všechny objekty (shluky) sloučeny do jediného shluku.

V případě divizního přístupu, jsou všechny objekty na počátku umístěny do jednoho shluku. Algoritmus DIANA tento shluk na základě zadaných kritérií (např. maximum euklidovské vzdálenosti mezi nejbližšími sousedy ve shluku) dělí na menší shluky. Tento proces dělení se opakuje tak dlouho, až nový shluk obsahuje pouze jediný objekt. [11]

Výhody hierarchického shlukování:

- Flexibilita týkající se úrovně nespojitosti
- Snadné zacházení se všemi formami podobnosti nebo vzdálenosti
- Aplikovatelnost na jakékoliv typy atributů

Nevýhody hierarchického shlukování:

- Vágnost cílového kritéria
- Skutečnost, že většina hierarchických algoritmů již zpětně nekontroluje jednou vytvořené shluky a tudíž je neoptimalizuje. [10]

3.2.1 Aglomerativní algoritmy

Při aglomerativním hierarchickém shlukování se vychází z matice vzdáleností vypočtených pro všechny páry objektů, popř. z jiné matice vztahů (obecně odlišností, včetně matice charakterizující vztahy proměnných či kategorií), což je jedna z jeho předností. Podle odlišných způsobů měření vzdáleností mezi shlukem C_g a sjednocením shluků C_h a $C_{h'}$, můžeme tento výpočet pro vybrané algoritmy zapsat následujícími způsoby [4], [5]:

- **metoda nejbližšího souseda** (jediné vazby, jednoduchého spojení (*single- link*)) - vzdálenost shluků je dána minimální vzdáleností objektů

  $$d_{g<h,h'>} = \frac{1}{2}\left(d_{gh} + d_{gh'} - \left|d_{gh} - d_{gh'}\right|\right) \quad (3.3)$$

- **metoda nejvzdálenějšího souseda** (úplné vazby, úplného spojení (*complete-link*)) – určující je maximální vzdálenost objektů

  $$d_{g<h,h'>} = \frac{1}{2}\left(d_{gh} + d_{gh'} + \left|d_{gh} - d_{gh'}\right|\right) \quad (3.4)$$

- **metoda průměrné vazby mezi shluky** (*average-link clustering*) – vzdálenost mezi dvěma shluky je počítána jako aritmetický průměr všech možných

vzdáleností objektů, z nichž jeden patří do prvního shluku a druhý do druhého shluku

$$d_{g<h,h'>} = \frac{n_h}{n_h + n_{h'}} d_{gh} + \frac{n_{h'}}{n_h + n_{h'}} d_{gh'} \quad (3.5)$$

- **Wardova metoda** – spojují se shluky, u nichž je přírůstek celkového vnitroskupinového součtu čtverců odchylek jednotlivých hodnot od shlukového průměru minimální

$$d_{g<h,h'>} = \frac{(n_h + n_g)d_{gh} + (n_{h'} + n_g)d_{gh'} - n_g d_{hh'}}{n_h + n_{h'} + n_g} \quad (3.6)$$

- **centroidní metoda** – vzdálenost mezi shluky je počítána jako euklidovská vzdálenost mezi jejich centroidy, což jsou vektory aritmetických průměrů počítané na základě všech objektů obsažených ve shluku

$$d_{g<h,h'>} = \frac{n_h}{n_h + n_{h'}} d_{gh} + \frac{n_{h'}}{n_h + n_{h'}} d_{gh'} - \frac{n_h n_{h'}}{(n_h + n_{h'})^2} d_{hh'} \quad (3.7)$$

- **mediánová metoda** – při výpočtu jsou na rozdíl od centroidní metody brány v úvahu velikosti shluků (počty jejich prvků)

$$d_{g<h,i>} = \frac{1}{2} d_{gh} + \frac{1}{2} d_{gh'} - \frac{1}{4} d_{hh'} \quad (3.8)$$

Vlastnosti vazebních metod:

- Nevýhodou metody nejbližšího souseda je tzv. efekt řetězení (*chaining effect*). Několik bodů, které tvoří most mezi dvěma shluky způsobuje, že metoda tyto dva shluky sjednocuje do jednoho.
- Shlukování metodou průměrné vazby může způsobovat slučování do podlouhlých shluků, čímž způsobuje částečný rozklad svých sousedů.
- Metody nejvzdálenějšího souseda obvykle vytváří více kompaktní a vhodnější hierarchii než metody nejbližšího souseda, ačkoliv ty jsou univerzálnější. [12]

Jednou z možností zlepšování kvality hierarchických metod je spojení hierarchického shlukování s dalšími shlukovacími technikami, v tomto případě mluvíme o vícefázovém shlukování. V literatuře je popsáno několik těchto specifických algoritmů, např.: BIRCH (*Balanced Iterative Reducing and Clustering Using Hierarchies*), ROCK (*RObust Clustering using linKs*), Chameleon, CURE (*Clustering Using REpresentatives*).

3.2.2 Divizní algoritmy

Při divizním shlukování se vychází ze stavu, kdy datové objekty tvoří jediný shluk, který se má rozdělit do dvou. Existují dva typy divizních algoritmů: monotetické

a polytetické. Mezi algoritmy využívající divizní shlukování patří např. BKMS (*Bisecting k-means*), DIANA (*DIvisive ANAlysis*) nebo DISMEA.

Monotetickou analýzou se označuje speciální typ divizního shlukování, aplikovaného na binární data. Rozdělení určitého shluku do dvou se provádí pouze podle jedné z proměnných (jedna skupina bude obsahovat v této proměnné jedničky, druhá nuly), což lze učinit podle libovolné proměnné. Při prvním dělení existuje p (počet proměnných) potencionálních rozdělení všech objektů do dvou skupin. Pro další dělení je k dispozici (p-1) možností atd. Kritérium pro dělení je založeno na měření závislosti dvou proměnných.

Monotetické shlukování má výhodu v tom, že po provedení analýzy lze nový objekt, který nebyl obsažen v původně analyzovaných datech, snadno přiřadit do některého z vytvořených shluků. Jelikož je každý shluk definován jedničkami nebo nulami určitých proměnných, jsou tím vytvořena alokační pravidla pro zařazení nových objektů.

Při polytetickém shlukování se hledají takové dva shluky, které mají nejmenší variabilitu z hlediska všech proměnných, jež charakterizují objekty. [5]

3.3 Metody založené na hustotě

Shluk je definován jako množina objektů spojených na základě hustoty, která je chápána ve smyslu četnosti a vzdálenosti objektů v určitých sousedstvích. Metody založené na hustotě (*Density-based methods*) jsou základem shlukování podprostorů. Cílem je nalézt podmnožiny proměnných tak, aby projekce datových objektů zahrnovaly oblasti s vysokou hustotou. Základem je rozdělení všech dimenzí do stejného počtu stejně dlouhých intervalů. Jsou-li určeny vhodné podprostory, spočívá úloha v nalezení shluků v odpovídajících projekcích. Shluky jsou oblasti navazujících jednotek s vysokou hustotou (v rámci určitého podprostoru). Shlukování podprostorů umožňuje zařazovat do shluků i těžko dosažitelné objekty. Výsledky jsou lepší než při nahrazení chybějících údajů hodnotami z příslušného rozdělení. [5]

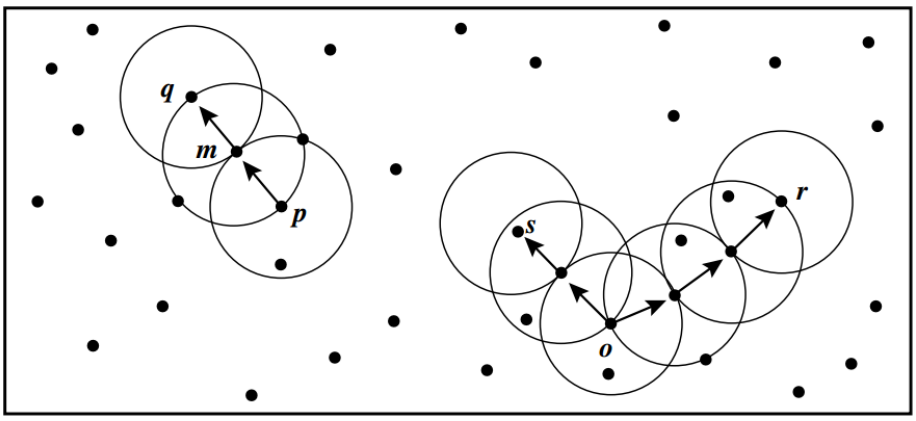
DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)

Algoritmus DBSCAN prohledává ε-sousedství (okolí objektu o poloměru ε) každého objektu v datové množině. Jestliže ε-sousedství bodu x obsahuje minimální počet objektů, tzv. *MinPts*, je vytvořen nový shluk s bodem x jako jádrem shluku. Algoritmus poté iterativně vyhledává na základě hustoty přímo dosažitelné objekty z těchto jader, což může vést ke spojování shluků. Proces končí, když již nelze žádný objekt přiřadit do shluku.

Dosažitelnost objektů:

- objekt x je na základě hustoty přímo dosažitelný z objektu y jestliže x se nachází v ε-sousedství y a y je jádrem shluku

- objekt *x* může být dosažitelný z objektu *y* na základě hustoty i za předpokladu, pokud existuje posloupnost objektů $x_1 \ldots x_n$, kde $x_1 = y$ a $x_n = x$, a když x_{t+1} je v sousedství x_t s ohledem na ε a *MinPts*



Obr. 3.4 DBSCAN [11]

Obrázek 3.4 ilustruje ε-sousedství objektu *m*, s *MinPts* = 3. Tento objekt je jádrem, neboť jeho sousedství obsahuje právě tři objekty. Každý z nich (např. *q*) je na základě hustoty přímo dosažitelný z objektu *m*. Objekty *p, o, r* jsou také jádry, jelikož jejich ε-sousedství obsahuje alespoň tři objekty. Objekt *q* je přímo dosažitelný z objektu *m*. Objekt *m* je naopak přímo dosažitelný z objektu *p*.

Objekt *q* je nepřímo dosažitelný z *p*, protože *q* je přímo dosažitelné z objektu *m* a *m* je přímo dosažitelný z *p*. Objekt *p* není dosažitelný z objektu *q*, jelikož *q* není jádrem. [11]

BRIDGE

Algoritmus BRIDGE využívá jednak metodu *k*-průměrů k rozkladu souboru dat na *k* shluky a současně algoritmus DBSCAN k nalezení hustoty shluků. [13]

DBCLASD (*Distribution-Based Clustering of LArge Spatial Databases*)

Algoritmus DBCLASD patří mezi tzv. smíšené metody, využívá principů metod založených na hustotě, mřížce a modelu. Identifikace shluků spočívá ve shlukování založeném na hustotě ve výběrovém prostoru (vzdálenost nejbližšího souseda pro objekty uvnitř oblasti jsou menší než pro objekty vně oblasti). K popisu shluku se využívá pravděpodobnostního rozdělení a zjišťuje se, zda po zahrnutí dalšího objektu do shluku je rozdělení stejné jako před přidáním objektu. K výhodám algoritmu DBCLASD patří schopnost nalezení shluků libovolného tvaru a určení počtu shluků bez potřeby zadávat parametry uživatelem. [4]

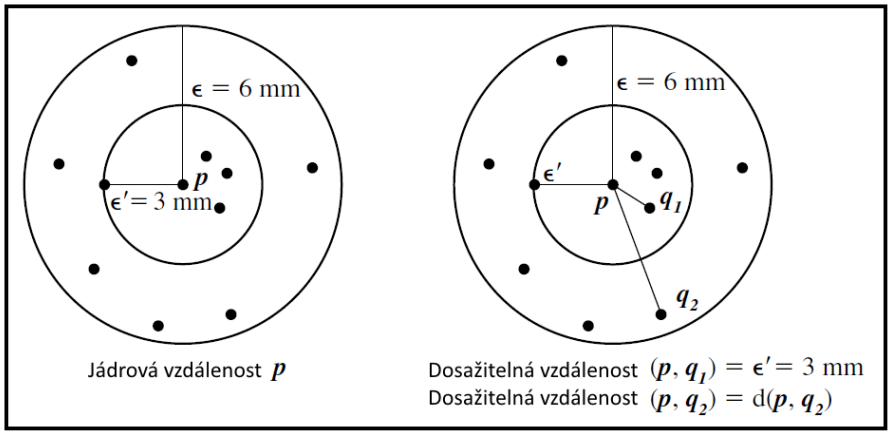
DENCLUE (*DENsity-Based CLUstering*)

Algoritmus DENCLUE je určen ke shlukování objemných multimediálních dat. Umí nalézt shluky libovolného tvaru a zároveň potlačit šum v datech. Algoritmus pracuje ve dvou krocích. Prvním krokem je vytvoření mřížky, která rozdělí data do hyperkvádrů (buněk). Do zpracování jsou zahrnuty pouze ty hyperkvádry, které obsahují objekty. Druhým krokem je vytvoření shluků. Shlukování se provádí pouze na základě vysoce obsazených hyperkvádrů a buněk, které jsou s nimi spojeny. [13]

OPTICS (*Ordering Points to Identify the Clustering Structure*)

Ačkoliv umí algoritmus DBSCAN shlukovat objekty se vstupními parametry ε a *MinPts*, stále nechává na uživateli odpovědnost za volbu dalších parametrů. Nastavení těchto parametrů je obvykle určeno empiricky a výpočet je obtížný, obzvláště pro vícedimenzionální množiny dat. Většina algoritmů je velmi citlivá na takové hodnoty parametrů, což může při jejich nesprávném nastavení vést k velmi rozdílným výsledkům shlukové analýzy.

Algoritmus OPTICS byl navržen k překonání těchto obtíží. Navrhuje uspořádání objektů v databázi. Pro objekty, které jsou jádrem, uchovává dvě hodnoty, a to jádrovou vzdálenost a dosažitelnou vzdálenost, aby mohly být objekty porovnávány (viz. obrázek 3.5). Tyto informace jsou postačující k získání takových objektů z databáze, které splňují podmínku vzdálenosti $\epsilon' < \epsilon$. [11]



Obr. 3.5 OPTICS [11]

Jádrová vzdálenost a dosažitelnost objektů:

- jádrová vzdálenost objektu p je nejmenší hodnota ε', která vytvoří z objektu {p} jádro. Jestliže p není jádrem, jádrová vzdálenost od objektu p není definována.
- dosažitelná vzdálenost objektu q od ostatních objektů p je dána hodnotou jádrové vzdálenosti p nebo euklidovskou vzdáleností mezi objekty p a q (větší z těchto dvou hodnot). Jestliže p není jádrem, dosažitelná vzdálenost mezi p a q není definována.

CUBN (*Clustering using Border and Nearest*)

Algoritmus CUBN využívá principů metod založených na hustotě a vzdálenosti. CUBN určí hraniční body pomocí eroze, což je jedna ze základních matematických morfologických operací, a poté shlukuje hraniční body a vnitřní body na základě výpočtu nejbližší vzdálenosti. Algoritmus umožňuje identifikovat nesférické shluky různé velikosti. Tento algoritmus se proto hodí i ke shlukování velkých datových souborů. [14]

V literatuře je zmíněno několik dalších algoritmů, které se využívají pro shlukování podprostorů: ENCLUS (*ENntrophy-Based CLUStering*), MAFIA (*Merging of Adaptive Finite Intervals (And More than a CLIQUE)*), OptiGrid (*Optimal Grid Clustering*), PROCLUS (*PROjected CLUStering*) nebo ORCLUS (*ORiented Projected CLUSter Generation*).

3.4 Metody založené na mřížce

Přístup metod založených na mřížce se liší od konvenčních shlukovacích algoritmů v tom, že nepracují s datovými body, ale zabývají se jejich okolím. Typický algoritmus shlukování založený na mřížce je obvykle tvořen pěti základními kroky [15]:

1) Navržení struktury mřížky, tj. rozdělení datového prostoru na konečný počet buněk.
2) Výpočet hustoty každé buňky.
3) Třídění buněk na základě jejich hustoty.
4) Identifikace středů shluků.
5) Propojení sousedních buněk.

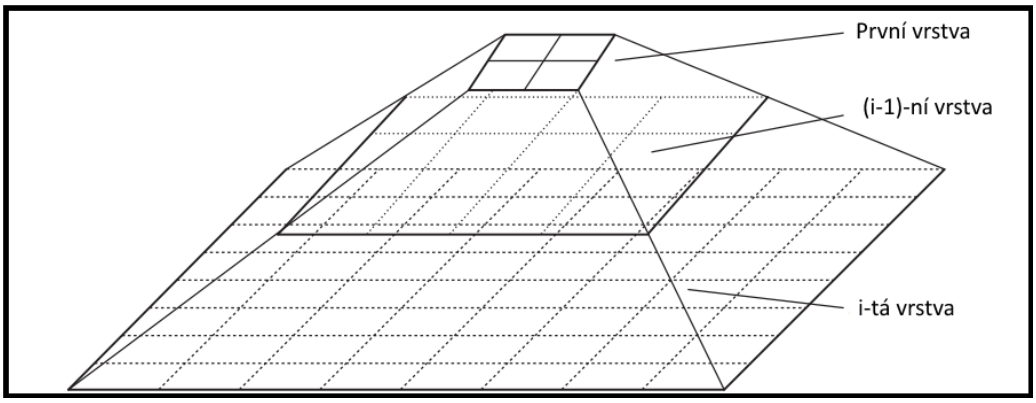
STING (*STatistical INformation Grid*)

Algoritmus STING dělí datový prostor do pravoúhlých buněk, které vytvářejí hierarchickou strukturu. Základem struktury je úroveň 1, na úrovni 2 jsou její potomci, atd. Hierarchická struktura shlukování založeném na algoritmu STING je zobrazena na obrázku 3.6. Tento algoritmus se vyznačuje výpočetní složitostí *O(K)*, kde *K* je počet buněk na úrovni 1. Z toho vyplývá, že použití této metody pro velký datový prostor je nevhodné. [13]

STING nabízí oproti ostatním shlukovacím metodám několik výhod:

- každá buňka obsahuje statistické informace potřebné pro výpočet velikosti základny hierarchické struktury
- mřížková struktura umožňuje paralelní zpracování a neustálou aktualizaci
- efektivita metody je její hlavní výhodou – STING prochází databází pouze jednou za účelem výpočtu statistických parametrů buněk, a proto je časová náročnost vytváření shluků *O(n)*, kde *n* je celkový počet objektů. Po vytvoření hierarchické struktury je výpočetní náročnost *O(g)*, kde *g* je celkový počet

pravoúhlých buněk na nejnižší úrovni struktury, přičemž obvykle platí, že *g* je mnohem menší než *n*. [1]



Obr. 3.6 STING [1]

CLIQUE (*CLustering In QUEst*)

CLIQUE je jednoduchá mřížková metoda pro nalezení shluků založených na hustotě v podprostoru. Metoda rozdělí každou dimenzi do navzájem nepřesahujících intervalů, čímž rozděluje celý vnitřní prostor datových objektů na buňky. K určení buněk s velkou hustotou a rozptýlených buněk využívá práh hustoty. Buňka má velkou hustotu, jestliže počet k ní náležících objektů překročí daný práh hustoty. [1]

WaveCluster

Tato technika shlukování definuje na datech homogenní dvourozměrnou mřížku a datové body jsou reprezentovány počtem bodů v každé buňce. Z datových bodů se stane množina bodů v odstínech šedi, s níž se dá zacházet jako s obrázkem. Tím se úloha hledání shluků transformuje na úlohu segmentaci obrazu, kde se vlnky s výhodou využívají pro svoje vlastnosti. [13]

V literatuře je zmíněno několik dalších metod založených na mřížce využívajících se ke shlukování dat: OptiGrid (*Optimal Grid Clustering*), GRIDCLUS nebo GDILC (*Grid-based Density-IsoLine Clustering*).

3.5 Metody založené na modelu

Zobecněním metody k-průměrů je shlukování založené na modelu (*Model-based methods*), v němž se předpokládá, že data lze vyjádřit modelem. Pokud bychom generovali data, pak si můžeme představit náhodný centroid, k němuž je přidán šum. Má-li tento normální rozdělení a kovariance je sférická, pak mají shluky kulový tvar. Shlukování založené na modelu předpokládá, že data vznikla výše popsaným způsobem. Cílem je odhadnout původní model, definovat shluky a přiřadit objekty ke

shlukům. K metodám založených na modelu a patří částicové filtry (*particle filters*) a algoritmus SOON (*Self Organizing Oscillator Network*).

Částicové filtry – využívá se filtr tvořený množinou částic a vah. Filtrovací metoda odhaduje množství důležitosti v množině N vážených částic. Pro případný nový objekt je množina sekvenčně aktualizována.

Algoritmus SOON – je založen na neuronové síti. Organizuje množinu objektů do *k* stabilních a strukturovaných shluků. Metoda vychází z algoritmu SOM (*Self-Organizing Map*), což je Kohonenova samoorganizující se neuronová síť. [4]

V literatuře se zmiňují další metody pro shlukování dat založené na modelu, a to COOLCAT a STUCCO.

4 EXPERIMENTÁLNÍ SROVNÁNÍ METOD SHLUKOVÉ ANALÝZY

Cílem této kapitoly je srovnání vybraných metod shlukové analýzy se zaměřením na úspěšnost při stanovení počtu shluků a zařazení jednotlivých instancí do správných tříd. Dalším cílem kapitoly je posouzení přínosu metody HFS pro selekci příznaků v úlohách bez učitele při shlukové analýze. Experimentální srovnání je provedeno na datech se známými výstupními třídami, tzv. *supervised* datech.

4.1 Vyvinutý software pro srovnání shlukovacích metod

Program pro experimentální srovnání metod shlukové analýzy byl vytvořen ve vývojovém prostředí MATLAB R2011 firmy *The MathWorks, Inc*. Byl využit *Statistic and Machine Learning Toolbox*, který poskytuje algoritmy a funkce pro analýzu a zpracování dat pomocí strojového učení.

Vyvinutý program představuje robustní nástroj pro selekci příznaků a srovnání metod shlukové analýzy. Program sestává z několika dílčích funkcí, které jsou volány prostřednictvím funkce `cluster_analysis`. Funkce provádí selekci příznaků metodou HFS, shlukovou analýzu předložených dat pomocí příslušného shlukovacího algoritmu, vyhodnocuje kvalitu výsledných shluků a vykresluje získané charakteristiky. Charakteristiky testovacích dat jsou podrobněji popsány v následující kapitole 4.2.

Implementace metody pro selekci příznaků HFS je volána funkcí `hfs`. Podrobný popis metody, algoritmu a použitých vzorců se nachází v kapitole 2.5.1. Kapitola 4.4 nabízí stručný popis vstupních parametrů funkce a příklad jejího volání. Pro ohodnocení významu příznaků (tzv. *rank*) lze využít dvou implementovaných algoritmů: lineárního korelačního koeficientu LCC a metody symetrické nejistoty SU.

Pro shlukování dat je možné využít algoritmy *k*-means (viz. kapitola 3.1.2), *single-link, complete-link, average-link* (viz. kapitola 3.2.1) a *Gaussian mixture model* s *EM* algoritmem. Kromě shlukovacích algoritmů se ve funkci `clusteval` volá externí funkce `cleval` z dostupného toolboxu LinkCluE [34], která poskytuje nástroje pro ohodnocení a porovnání kvality výsledných shluků využitím interních a externích validačních kritérií. Jsou implementována tři interní (DB, CP, Dunn) a tři externí (CA, RI, AR) validační kritéria, jejichž podrobný popis je uveden v kapitole 2.4.1.

Funkci `cluster_analysis` lze volat bez vstupních parametrů nebo s nepovinnými doplňkovými vstupními parametry. Při volání funkce bez parametrů se provede vyhodnocení defaultního datového setu *Fisher´s Iris* metodou *k*-means a interními validačními indexy. Nepovinné vstupní parametry jsou: *sampledata, clustmethod, truelabels, hfs* a *threshold*. První parametr označuje jméno datového setu, druhý použitý shlukovací algoritmus. Pokud není parametr *truelabels* prázdný, jsou k dispozici známé třídy dat a kvalita výsledných shluků je ohodnocena interními i externími validačními

kritérii. Dalšími volitelnými parametry jsou *hfs* a *threshold*. Pokud není parametr *hfs* prázdný, jsou data předzpracována metodou selekce příznaků HFS se zvoleným prahem.

Příklad volání funkce:

```
> cluster_analysis('ecoli', 'clink', 1, 1, 0.65)
```

Voláním funkce `cluster_analysis` s výše uvedenými parametry budou data v datovém setu *E.coli* redukována metodou pro selekci příznaků HFS s prahem 0,65. Bude použit shlukovací algoritmus *complete-link* a kvalita výsledných shluků bude ohodnocena interními i externími validačními kritérii. Podrobný popis jednotlivých funkcí a nastavení jejich parametrů při volání se nachází v hlavičce každé funkce zdrojového programu. V prostředí Matlab je nápověda přístupná po zavolání příkazu `help <jmeno_funkce>` v adresáři s programem.

4.2 Charakteristika datových setů

K srovnání metod shlukové analýzy a pro posouzení přínosu metody selekce příznaků HFS pro úspěšnost shlukové analýzy jsou vybrány čtyři reálné datové soubory z UCI repozitáře pro strojové učení [35] a datový soubor naměřených reálných dat na vibračním přípravku.

a) Datové sety z UCI repozitáře

UCI repozitář poskytuje velké množství datových setů se známými třídami použitelných pro strojové učení. Většina z těchto datových setů se běžně využívá k testování výkonnosti shlukovacích algoritmů a metod pro selekci příznaků. Detailní informace o vybraných šesti datových setech jsou přehledně zobrazeny v tabulce 4.1.

Fisher´s Iris

Tento datový set je patrně nejznámější databází využívanou pro strojové učení. Set obsahuje 150 instancí charakterizovaných čtyřmi příznaky a klasifikovaných do tří tříd. Každá výstupní třída odpovídá jednomu typu kosatce (*iris*). Výstupní třídy jsou: *Iris Setosa*, *Iris Versicolor* a *Iris Virginica*. Třída *Iris Setosa* je lineárně separovatelná od ostatních dvou tříd.

E.coli

Datový set *E.coli* je soubor nalezených umístění 336 proteinů *E.coli* klasifikující je do osmi tříd: cytoplasma (cp), vnitřní membrána bez signálové sekvence (im), periplasma (pp), vnitřní membrána s nerozštěpenou sekvencí (imU), vnější membrána (om), vnější membrána s lipoproteinem (omL), vnitřní membrána s lipoproteinem (imL)

a vnitřní membrána s rozštěpenou sekvencí (imS). Každý protein je popsán sedmi příznaky: mcg, gvh, lip, chg, aac, alm1 a alm2.

WDBC

Datový set WDBC (*Wisconsin Diagnostic Breast Cancer*) byl poprvé použit ke zkoumání maligní a benigní formy rakoviny prsu. Obsahuje příznaky získané z digitalizovaného obrázku prsní tkáně získané velmi tenkou jehlou (FNA, f*ine needle aspirate*). Příznaky popisují charakteristiky jádra buňky přítomné na získaném obrázku. Datový set obsahuje 569 instancí charakterizovaných 32 příznaky a klasifikovaných do dvou tříd.

LSTV

Datový set *LSVT Voice Rehabilitation* se sestává z biometrických algoritmů pro zpracování řeči. Bylo použito 310 algoritmů k popsání 126 řečových signálů získaných od 14 osob, u kterých byla diagnostikována Parkinsonova choroba. Výstupními třídami jsou fonační reakce při rehabilitaci Parkinsonovy choroby: *Acceptable* a *Unacceptable*.

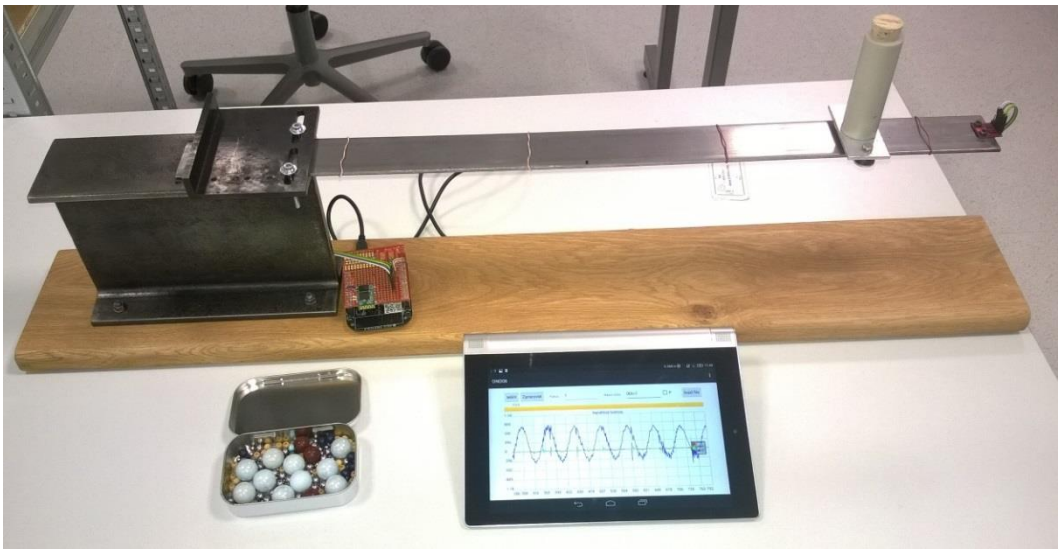
	Počet instancí	Počet příznaků	Počet tříd
Fisher´s Iris	150	4	3
E.coli	336	7	8
Wisconsin Diagnostic Breast Cancer	569	30	2
LSVT Rehabilitation [37]	126	310	2

Tab. 4.1 Charakteristiky datových setů

b) Data z vibračního přípravku

Data byla naměřena na vibračním přípravku (obrázek 4.1) vyhotoveném v rámci diplomové práce „Sběr dat a detekce anomálií přes mobilní zařízení“ Bc. Michaela Ondráška mobilním zařízením s OS Android. Aplikace na mobilním zařízení komunikuje s komunikačním modulem na přípravku s použitím bezdrátové technologie *Bluetooth*. Data se přes datové spojení posílají na server, kde probíhá jejich zpracování, vyhodnocení a archivace.

Na vibračním přípravku se měří časové průběhy s proměnným počtem a typem závaží. Závaží tvoří velké ložiskové kuličky a skleněné kuličky, které se vkládají do tubusu. Měření časových průběhů byla provedena bez závaží, s jednou/dvěma/pěti velkými ložiskovými kuličkami a s jednou nebo dvěma skleněnými kuličkami. Pro každé měření bylo vypočteno 17 příznaků pro každou osu, tj. celkem 51 příznaků. Souhrnné informace o naměřeném datovém setu jsou uvedeny v tabulce 4.2.

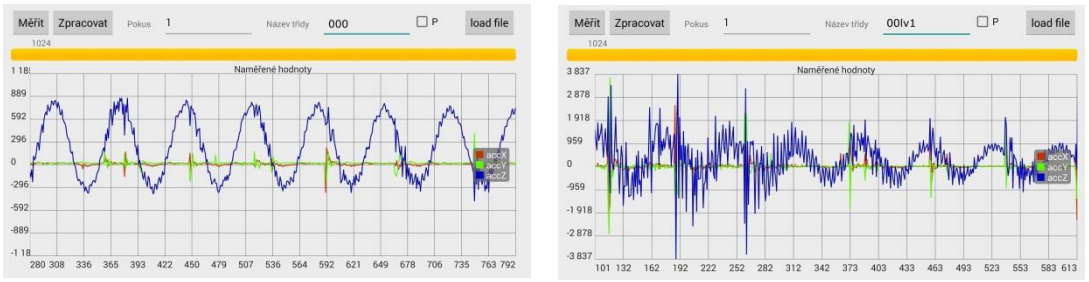


Obr. 4.1 Vibrační přípravek

	Počet instancí	Počet příznaků	Počet tříd
Set naměřených dat	181	51	6

Tab. 4.2 Informace o souboru naměřených dat

Obrázek 4.2 vlevo zobrazuje naměřený průběh bez závaží a na vpravo je zobrazen časový průběh s jednou velkou ložiskovou kuličkou.



Obr. 4.2 Časové průběhy bez závaží (vlevo) a s jednou ložiskovou kuličkou (vpravo)

4.3 Stanovení počtu shluků a zařazení instancí do tříd

V předcházející třetí kapitole bylo přestaveno množství shlukovacích algoritmů určených pro různé aplikace a datové struktury. Žádný shlukovací algoritmus ale není vhodný pro použití se všemi datovými strukturami. Shlukování je proces učení bez učitele (*unsupervised learning*), kde předem není znám správný počet výstupních tříd a nevíme, zda nalezené shluky jsou validní. Aby bylo možné mezi sebou porovnat výsledky shlukování různými shlukovacími algoritmy, je potřeba použít validačních kritérií. Jejich základní přehled je uveden v kapitole 2.4.

V této kapitole jsou porovnávány výsledky shlukování pomocí algoritmů závislých na správném počátečním nastavení počtu shluků (*k*-means, *single-link, complete-link, average-link* a *EM*). Iterativně se nastavuje počet shluků v rozsahu od dvou do *k,* a výsledky shlukování jsou ohodnoceny interními (DB, CP, Dunn) a externími (CA, RI, AR) validačními kritérii. Na jejich základě je určen optimální počet shluků pro každý datový set. Připomeňme, že nízké hodnoty validačních indexů CP a DB značí dobrou strukturu shluků, naopak u validačních indexů Dunn, CA, RI a AR ukazují dobrou kvalitu shluků jejich vysoké hodnoty. Testování bylo provedeno na vybraných datových setech (viz. kapitola 4.2) se známými výstupními třídami.

Jelikož jsou známy výstupní třídy vybraných datových setů, je možné pro jednotlivé výsledky shlukování určit správnost zařazení instancí do tříd. Pro každý datový set a shlukovací algoritmus jsou chybně klasifikované instance vyjádřeny graficky a v tabulce jsou uvedeny počty chybně klasifikovaných instancí do tříd.

4.3.1 Fisher´s Iris

4.3.1.1 Stanovení počtu shluků

V tabulkách 4.3 až 4.7 jsou uvedeny hodnoty interních a externích validačních kritérií získaných při evaluaci výsledků shlukování algoritmy *k*-means, *single-link, complete-link, average-link* a *EM* algoritmu při použití datového setu Fisher´s Iris. Zvýrazněny jsou minimální/maximální hodnoty validačních indexů určující stanovený počet shluků.

Val. index	2	3	4	5	6	7	8	9	10
CP	1,217	0,919	0,863	0,749	0,696	0,637	0,607	**0,578**	0,615
DB	**0,404**	0,589	0,691	0,736	0,840	0,763	0,830	0,753	0,837
Dunn	**3,915**	2,435	1,077	1,171	1,170	1,302	1,046	0,944	0,877
AR	0,540	**0,730**	0,602	0,693	0,537	0,418	0,430	0,414	0,552
RI	0,764	**0,880**	0,830	0,876	0,821	0,776	0,787	0,782	0,827
CA	0,667	0,893	0,893	**0,980**	0,973	0,907	0,967	0,967	0,947

Tab. 4.3 Validační indexy, datový set Fisher´s Iris (*k*-means)

Val. index	2	3	4	5	6	7	8	9	10
CP	1,211	1,166	1,153	1,076	1,058	1,045	1,036	1,028	**1,009**
DB	0,383	0,430	0,237	0,296	0,225	0,199	0,165	0,153	**0,144**
Dunn	**3,824**	2,502	2,500	2,384	0,000	0,000	0,000	0,000	0,000
AR	**0,568**	0,564	0,562	0,552	0,550	0,540	0,537	0,535	0,533
RI	0,776	**0,777**	0,777	0,777	0,777	0,773	0,773	0,773	0,773
CA	0,667	0,680	0,687	0,693	0,693	0,693	0,707	0,707	**0,713**

Tab. 4.4 Validační indexy, datový set Fisher´s Iris (*single-link*)

Val. index	2	3	4	5	6	7	8	9	10
CP	1,487	0,954	0,827	0,775	0,679	0,646	0,609	0,598	**0,584**
DB	0,657	**0,537**	0,638	0,719	0,711	0,709	0,742	0,654	0,690
Dunn	**2,452**	2,097	2,218	1,094	1,266	1,266	1,163	1,163	1,196
AR	0,422	**0,642**	0,589	0,445	0,449	0,424	0,457	0,462	0,452
RI	0,711	**0,837**	0,822	0,769	0,785	0,777	0,796	0,798	0,795
CA	0,660	0,840	0,840	0,840	0,893	0,893	0,967	0,973	**0,973**

Tab. 4.5 Validační indexy, datový set Fisher´s Iris (*complete-link*)

Val. index	2	3	4	5	6	7	8	9	10
CP	1,211	0,921	0,865	0,796	0,783	0,688	0,676	0,650	**0,640**
DB	**0,383**	0,585	0,530	0,565	0,468	0,504	0,575	0,588	0,570
Dunn	**3,824**	2,420	2,390	1,781	1,781	1,663	1,598	1,598	0,000
AR	0,568	**0,759**	0,729	0,664	0,652	0,579	0,572	0,525	0,530
RI	0,776	**0,892**	0,881	0,855	0,851	0,830	0,828	0,811	0,813
CA	0,667	0,907	0,907	0,907	0,907	0,907	0,907	0,907	**0,913**

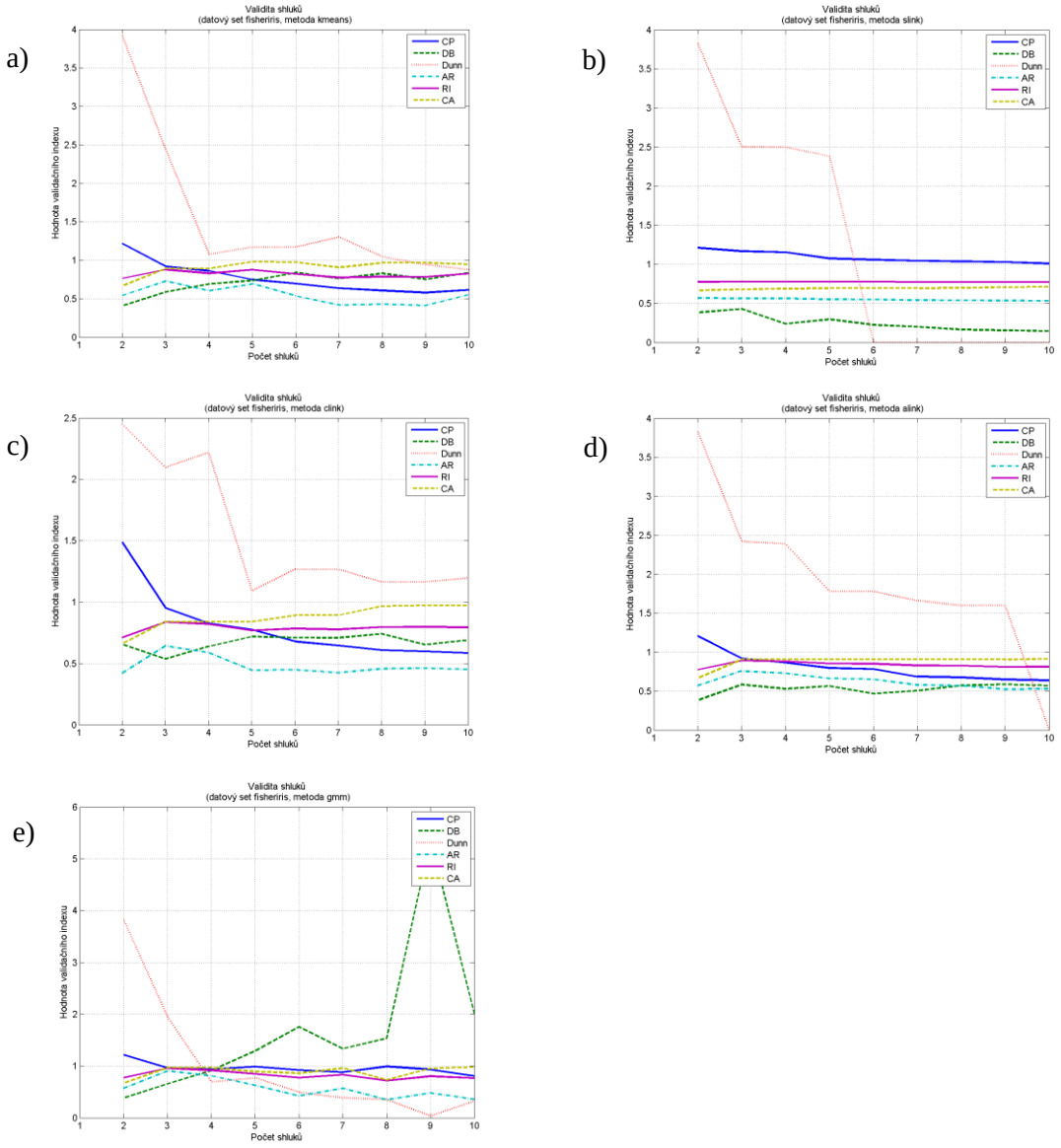
Tab. 4.6 Validační indexy, datový set Fisher´s Iris (*average-link*)

Val. index	2	3	4	5	6	7	8	9	10
CP	1,211	0,958	0,939	0,986	0,923	0,881	0,992	0,934	**0,812**
DB	**0,383**	0,654	0,912	1,292	1,756	1,333	1,533	5,355	2,004
Dunn	**3,824**	1,948	0,694	0,772	0,494	0,386	0,353	0,039	0,321
AR	0,568	**0,904**	0,811	0,626	0,423	0,572	0,350	0,483	0,356
RI	0,776	**0,957**	0,919	0,848	0,775	0,832	0,719	0,804	0,766
CA	0,667	0,967	0,967	0,893	0,860	0,960	0,733	0,953	**0,980**

Tab. 4.7 Validační indexy, datový set Fisher´s Iris (*EM*)

Obrázek 4.3 graficky zobrazuje hodnoty validačních indexů, které jsou uvedeny v tabulkách 4.3 až 4.7. Největší počet správných shluků v datovém setu Fisher´s Iris bylo určeno pomocí externích validačních kritérií AR a RI. Pouze v jednom případě, při

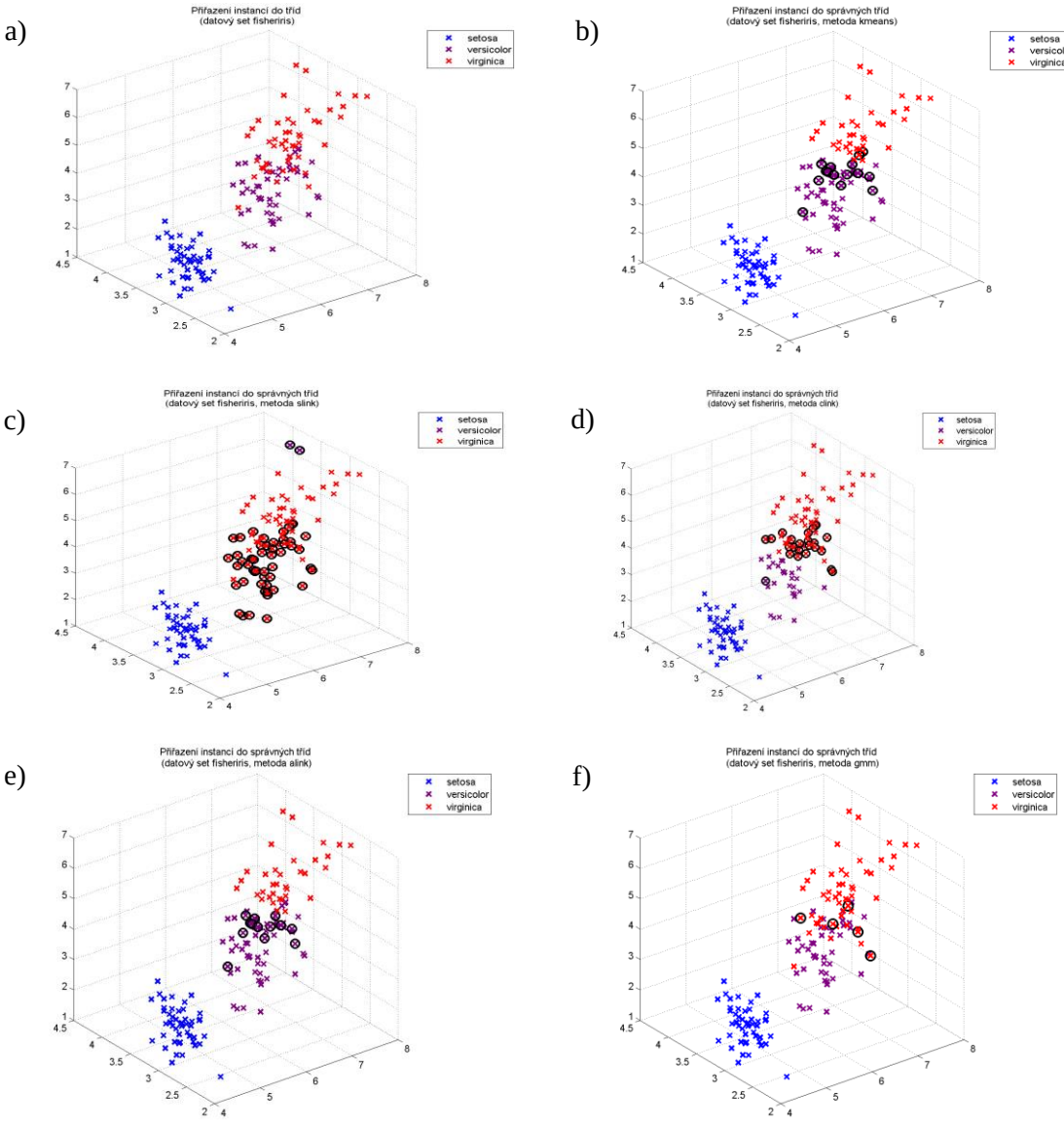
shlukování dat pomocí algoritmu *single-link*, nedošlo ke správnému vyhodnocení počtu shluků validačním indexem AR. Validační kritérium DB bylo schopno určit dobrou strukturu shluků pouze při použití algoritmu *complete-link*. Ostatní validační kritéria nebyla schopna určit při daném výsledku shlukování správný počet shluků předložených dat. Příčinou je, že shluky nejsou od sebe v datovém setu dobře separovány. V několika případech kritéria DB a Dunn ohodnotila výsledek shlukování tak, že bylo vyhodnoceno, že data jsou rozdělena pouze do dvou tříd.



Obr. 4.3 Fisher´s Iris - validita shluků, a) *k*-means, b) single-link, c) complete-link, d) average-link, e) EM

4.3.1.2 Přiřazení instancí do tříd

Obrázek 4.4 zobrazuje grafické přiřazení instancí do tříd použitého datového setu a přiřazení instancí do tříd po použití jednotlivých shlukovacích algoritmů s vyznačením chybně klasifikovaných bodů černým kroužkem.



Obr. 4.4 Fisher´s Iris - přiřazení instancí do tříd, a) vstupní data b) *k*-means, c) single-link, d) complete-link, e) average-link, f) EM

Tabulka 4.8 uvádí počty instancí v jednotlivých třídách datového setu Fisher´s Iris a zařazení instancí do tříd ve výsledcích shlukování danými shlukovacími algoritmy.

Třída	**Setosa**	**Versicolor**	**Virginica**
# instancí	**50**	**50**	**50**
k-means	50	0	0
	0	48	14
	0	2	36
single-link	50	0	0
	0	0	1
	0	50	48
complete-link	50	0	0
	0	27	1
	0	23	49
average-link	50	0	0
	0	50	14
	0	0	36
EM	50	0	0
	0	45	0
	0	5	50

Tab. 4.8 Fisher´s Iris - přiřazení instancí do tříd

V tabulce 4.9 jsou celkové součty chybně klasifikovaných přiřazení instancí do tříd a jejich procentuální vyjádření.

	k-means	single-link	complete-link	average-link	EM
Chybně klasifikováno	16	52	24	14	5
Vyjádření v [%]	10,6	34,6	16,0	9,3	3,3

Tab. 4.9 Fisher´s Iris – chybně přiřazené instance do tříd

Z grafického vyjádření na obrázku 4.4 i číselného vyjádření v tabulkách 4.8 a 4.9 je patrné, že *EM* algoritmus nepřiřadil pouze pět instancí (3,3 %) do správné třídy. Další testované shlukovací algoritmy *k*-means, *average-link* a *complete-link* jsou ale také schopny úspěšně zařadit velké procento instancí datového setu Fisher´s Iris do jednotlivých tříd. Algoritmus *single-link* nebyl schopen rozlišit málo separované třídy *Versicolor* a *Virginica* a chybně klasifikoval 34,6 % instancí.

4.3.2 E.coli

4.3.2.1 Stanovení počtu shluků

V tabulkách 4.10 až 4.14 jsou uvedeny hodnoty interních a externích validačních kritérií získaných při evaluaci výsledků shlukování algoritmy *k*-means, *single-link*, *complete-link*, *average-link* a *EM* algoritmu při použití datového setu E.coli. Zvýrazněny jsou minimální/maximální hodnoty validačních indexů určující stanovený počet shluků.

Val. index	2	3	4	5	6	7	8	9	10
CP	0,412	0,341	0,323	0,312	0,293	0,285	0,272	0,268	0,253
DB	0,976	0,985	1,111	1,232	1,072	1,132	1,071	1,149	**0,965**
Dunn	**1,835**	1,671	1,066	1,061	0,851	0,830	0,572	0,621	0,591
AR	0,379	**0,667**	0,542	0,452	0,372	0,380	0,510	0,382	0,353
RI	0,671	**0,856**	0,820	0,795	0,781	0,786	0,830	0,794	0,790
CA	0,622	0,750	0,744	0,717	0,765	0,795	0,818	0,827	0,830

Val. index	11	12	13	14	15	16	17	18	
CP	0,248	0,243	0,240	0,233	0,233	0,231	**0,222**	0,223	
DB	1,135	1,058	1,127	1,171	1,089	1,115	1,110	1,037	
Dunn	0,509	0,645	0,910	0,536	0,575	0,651	0,672	0,493	
AR	0,356	0,322	0,330	0,282	0,277	0,300	0,254	**0,252**	
RI	0,795	0,784	0,789	0,780	0,778	0,785	0,774	0,774	
CA	**0,848**	0,839	0,866	0,863	0,848	0,875	0,884	0,878	

Tab. 4.10 Validační indexy, datový set E.coli (*k*-means)

Val. index	2	3	4	5	6	7	8	9	10
CP	0,541	0,527	0,523	0,520	0,519	0,516	0,513	0,511	0,508
DB	0,299	0,711	0,707	0,553	0,488	0,275	0,245	0,222	0,204
Dunn	**3,344**	1,469	1,400	0,000	0,000	0,000	0,000	0,000	0,000
AR	0,004	0,038	0,038	0,038	0,038	0,039	0,040	0,041	0,042
RI	0,276	0,324	0,324	0,324	0,324	0,327	0,330	0,333	0,336
CA	0,429	0,443	0,446	0,449	0,449	0,452	0,455	0,458	0,461

Val. index	11	12	13	14	15	16	17	18	
CP	0,506	0,504	0,501	0,499	0,496	0,495	0,492	**0,489**	
DB	0,190	0,179	0,233	0,199	0,189	0,182	0,174	**0,168**	
Dunn	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	
AR	0,043	0,046	0,048	0,045	0,048	0,049	0,050	**0,054**	

RI	0,340	0,344	0,351	0,351	0,356	0,359	0,362	**0,367**	
CA	0,464	0,467	0,473	0,473	0,476	0,479	0,482	**0,485**	

Tab. 4.11 Validační indexy, datový set E.coli (*single-link*)

Val. index	2	3	4	5	6	7	8	9	10
CP	0,412	0,343	0,337	0,327	0,320	0,316	0,309	0,293	0,281
DB	0,969	0,907	1,141	1,101	1,120	0,921	0,971	0,950	0,967
Dunn	**1,836**	1,676	1,157	0,838	0,838	1,264	1,050	1,050	1,050
AR	0,386	**0,692**	0,664	0,684	0,690	0,691	0,616	0,593	0,622
RI	0,674	0,867	0,858	0,868	**0,872**	**0,872**	0,846	0,843	0,857
CA	0,625	0,759	0,759	0,774	0,774	0,777	0,777	0,777	0,824

Val. index	11	12	13	14	15	16	17	18	
CP	0,271	0,268	0,258	0,251	0,249	0,246	0,242	**0,238**	
DB	1,037	1,055	0,998	1,025	0,933	**0,881**	0,941	0,949	
Dunn	0,758	0,741	0,726	0,726	0,726	0,000	0,000	0,000	
AR	0,448	0,449	0,367	0,352	0,352	0,352	0,346	0,353	
RI	0,809	0,809	0,788	0,790	0,790	0,790	0,791	0,794	
CA	0,824	0,833	0,833	0,842	0,845	0,848	0,863	**0,875**	

Tab. 4.12 Validační indexy, datový set E.coli (*complete-link*)

Val. index	2	3	4	5	6	7	8	9	10
CP	0,532	0,400	0,397	0,386	0,383	0,378	0,308	0,305	0,303
DB	1,056	0,916	0,808	0,867	0,772	0,816	0,767	0,658	0,628
Dunn	1,559	**1,895**	1,799	1,438	1,438	1,438	1,444	1,837	0,000
AR	0,030	0,401	0,412	0,461	0,462	0,463	**0,745**	0,744	0,744
RI	0,310	0,687	0,694	0,731	0,731	0,733	**0,894**	**0,894**	**0,894**
CA	0,440	0,631	0,634	0,661	0,664	0,664	0,798	0,798	0,801

Val. index	11	12	13	14	15	16	17	18	
CP	0,298	0,296	0,294	0,283	0,271	0,268	0,267	**0,266**	
DB	0,648	0,623	0,635	0,647	0,641	0,631	**0,588**	0,596	
Dunn	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	
AR	0,706	0,706	0,706	0,691	0,716	0,696	0,696	0,692	
RI	0,880	0,880	0,880	0,878	0,890	0,884	0,884	0,882	
CA	0,801	0,801	0,801	0,801	**0,842**	**0,842**	**0,842**	**0,842**	

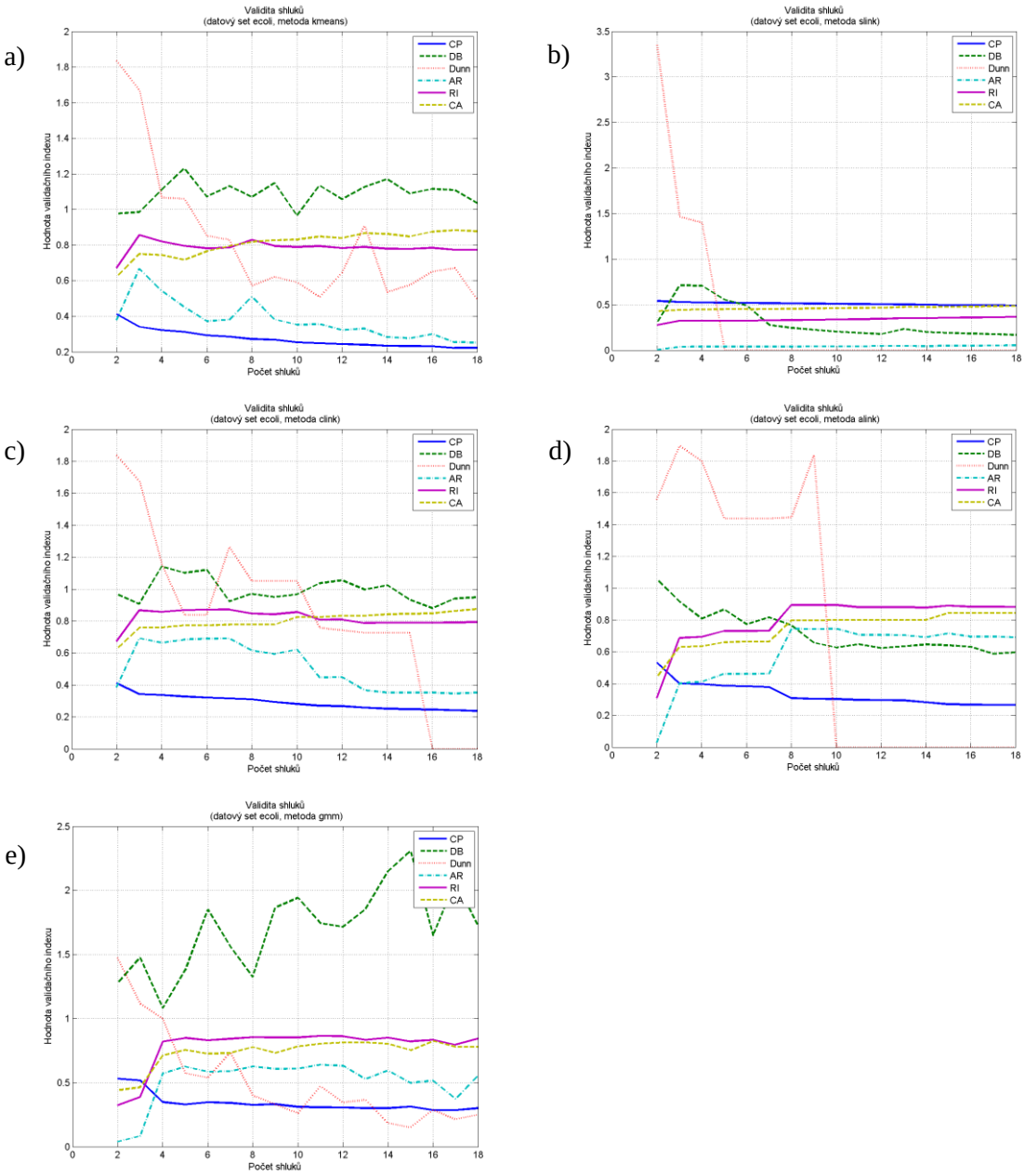
Tab. 4.13 Validační indexy, datový set E.coli (*average-link*)

Val. index	2	3	4	5	6	7	8	9	10
CP	0,530	0,517	0,348	0,330	0,347	0,341	0,326	0,331	0,311
DB	1,274	1,478	**1,081**	1,376	1,850	1,562	1,324	1,868	1,942
Dunn	**1,472**	1,115	0,998	0,573	0,538	0,737	0,399	0,331	0,260
AR	0,038	0,082	0,572	0,624	0,585	0,588	0,626	0,606	0,609
RI	0,323	0,387	0,822	0,849	0,829	0,841	0,855	0,852	0,852
CA	0,440	0,464	0,711	0,756	0,726	0,729	0,777	0,732	0,783

Val. index	11	12	13	14	15	16	17	18	
CP	0,308	0,305	0,301	0,301	0,313	**0,285**	0,286	0,300	
DB	1,744	1,715	1,853	2,151	2,308	1,648	2,074	1,725	
Dunn	0,472	0,345	0,366	0,183	0,151	0,287	0,212	0,250	
AR	**0,640**	0,632	0,528	0,594	0,498	0,516	0,372	0,554	
RI	**0,865**	0,863	0,833	0,851	0,822	0,834	0,794	0,842	
CA	0,804	0,813	0,813	0,804	0,753	**0,824**	0,780	0,780	

Tab. 4.14 Validační indexy, datový set E.coli (*EM*)

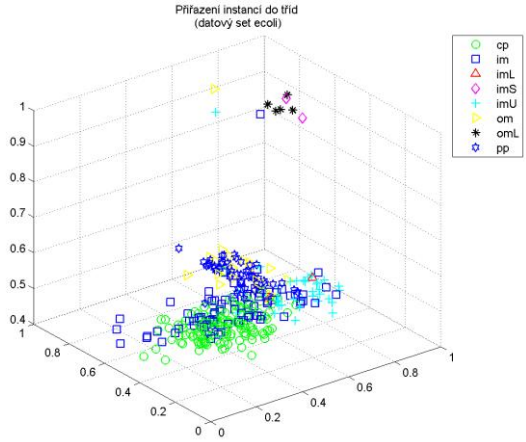
Obrázek 4.5 graficky zobrazuje hodnoty validačních indexů, které jsou uvedeny v tabulkách 4.10 až 4.14. Přesný počet osmi shluků byl určen pouze při použití shlukovacího algoritmu *average-link* externími kritérii AR a RI. V ostatních případech nebyl ani jednou určen správný počet shluků datového setu validačními kritérii. Shluky jsou v tomto setu ještě méně separovány, než u setu *Fisher´s Iris* a pro použité shlukovací algoritmy je obtížné vytvořit odpovídající shlukovou strukturu.



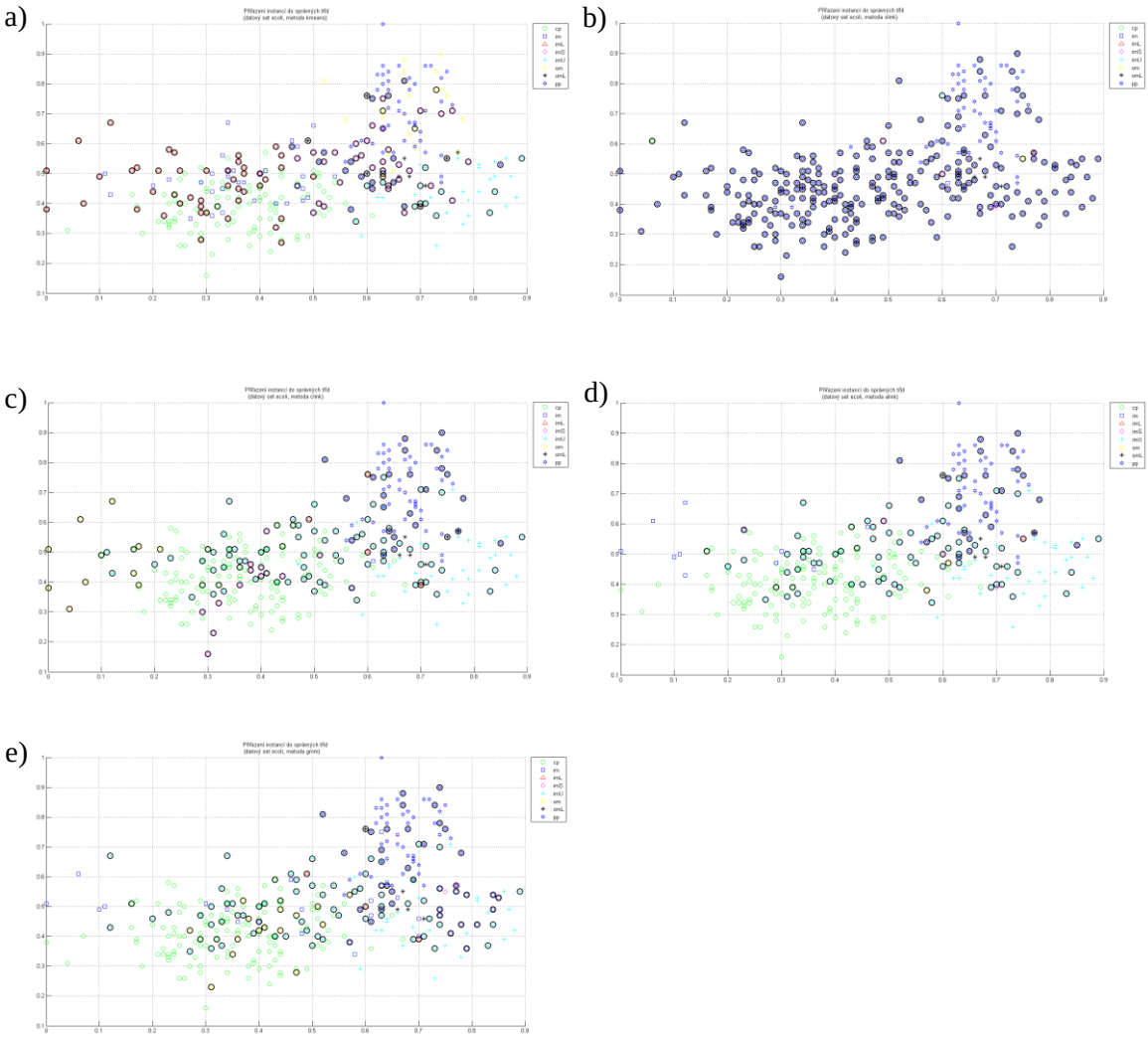
Obr. 4.5 E.coli - validita shluků, a) *k*-means, b) single-link, c) complete-link, d) average-link, e) EM

4.3.2.2 Přiřazení instancí do tříd

Obrázek 4.6 zobrazuje grafické přiřazení instancí do tříd použitého datového setu v prostorovém zobrazení. Pro přehlednost jsou na obrázku 4.7 zobrazena přiřazení instancí do shluků po použití jednotlivých shlukovacích algoritmů v jedné rovině. Chybně klasifikované body jsou vyznačeny černým kroužkem.



Obr. 4.6 E.coli - přiřazení instancí do tříd



Obr. 4.7 E.coli - přiřazení instancí do správných tříd, a) *k*-means, b) single-link, c) complete-link, d) average-link, e) EM

Tabulka 4.15 uvádí počty instancí v jednotlivých třídách datového setu E.coli a zařazení instancí do tříd ve výsledcích shlukování danými shlukovacími algoritmy.

Třída	**cp**	**im**	**pp**	**imU**	**om**	**omL**	**imL**	**imS**
# instancí	**143**	**77**	**2**	**2**	**35**	**20**	**5**	**52**
k-means	97	0	0	0	0	0	0	1
	0	31	0	0	1	0	0	0
	36	8	0	0	0	0	0	2
	0	21	1	0	9	0	0	1
	0	14	0	0	23	0	0	0
	0	0	0	0	0	18	0	4
	0	1	0	2	1	1	5	0
	10	2	1	0	1	1	0	44
single-link	0	1	0	0	0	0	0	0
	0	1	0	0	0	0	0	0
	0	0	0	0	0	0	1	0
	0	1	0	1	1	0	0	0
	0	0	0	0	0	1	0	0
	0	0	0	1	0	0	0	0
	0	0	0	0	0	0	4	0
	143	74	2	0	34	19	0	52
complete-link	120	6	0	0	0	0	0	3
	0	1	0	0	1	0	0	0
	0	1	0	1	1	1	0	0
	13	0	0	0	0	0	0	0
	0	66	1	0	33	0	0	1
	7	3	0	0	0	0	0	0
	0	0	0	1	0	0	5	0
	3	0	1	0	0	19	0	48
average-link	139	2	0	0	0	0	0	3
	1	10	0	0	0	0	0	0
	0	0	0	1	0	0	0	0
	0	1	0	1	1	0	0	0
	0	63	1	0	33	0	0	1
	0	1	0	0	1	0	0	0
	0	0	0	0	0	1	5	0
	3	0	1	0	0	19	0	48
EM	123	2	0	0	0	0	0	4
	1	18	1	0	9	0	0	4
	0	1	0	1	1	0	0	0

	0	0	0	1	0	0	1	0
	0	56	1	0	25	0	0	1
	15	0	0	0	0	0	0	0
	0	0	0	0	0	1	4	0
	4	0	0	0	0	19	0	43

Tab. 4.15 E.coli - přiřazení instancí do tříd

V tabulce 4.16 jsou celkové součty chybně klasifikovaných přiřazení instancí do tříd a jejich procentuální vyjádření.

	k-means	single-link	complete-link	average-link	EM
Chybně klasifikováno	118	278	129	100	122
Vyjádření v [%]	35,1	82,7	38,4	29,8	36,3

Tab. 4.16 E.coli – chybně přiřazené instance do tříd

Z grafického vyjádření na obrázku 4.7 i číselného vyjádření v tabulkách 4.15 a 4.16 je patrné, že žádný z testovaných algoritmů nebyl schopen rozpoznat a přiřadit všechny instance do správných tříd. Nejlepší výsledek vykazuje algoritmus *average-link*, ale i ten nepřiřadil správně téměř 30% všech instancí. Srovnatelné výsledky mají algoritmy *k*-means, *EM* a *complete-link*, jež mají chybovost přiřazení instancí 35,1 %, 36,3 % a 38,4 %. Algoritmus *single-link* není pro málo separované shluky instancí vůbec vhodný, nepřiřadil správně přes 80 % instancí do tříd.

4.3.3 WDBC

Datový set WDBC (*Wisconsin Diagnostic Breast Cancer*) byl poprvé použit ke zjištění maligní a benigní formy rakoviny prsu. Tento datový set obsahuje příznaky získané z digitalizovaného obrázku prsní tkáně odebrané velmi tenkou jehlou (FNA, f*ine needle aspirate*). Příznaky popisují charakteristiky jádra buňky přítomné na získaném obrázku. Datový set obsahuje 569 instancí charakterizovaných 32 příznaky a klasifikovaných do dvou tříd.

4.3.3.1 Stanovení počtu shluků

V tabulkách 4.17 až 4.21 jsou uvedeny hodnoty interních a externích validačních kritérií získaných při evaluaci výsledků shlukování algoritmy *k*-means, *single-link, complete-link, average-link* a *EM* algoritmem při použití datového setu WDBC. Zvýrazněny jsou minimální/maximální hodnoty validačních indexů určující stanovený počet shluků.

Val. index	2	3	4	5	6	7	8	9	10
CP	378,7	295,7	241,8	204,6	175,5	159,5	142,9	132,1	**126,9**
DB	0,504	0,626	0,574	0,526	**0,477**	0,527	0,594	0,538	0,662
Dunn	**2,889**	1,353	0,861	0,730	0,520	0,487	0,414	0,303	0,355
AR	0,491	**0,541**	0,413	0,351	0,268	0,235	0,200	0,169	0,183
RI	0,750	**0,769**	0,701	0,668	0,621	0,604	0,584	0,566	0,573
CA	0,854	0,889	0,835	0,886	0,875	0,888	0,891	0,909	**0,910**

Tab. 4.17 Validační indexy, datový set WDBC (*k*-means)

Val. index	2	3	4	5	6	7	8	9	10
CP	673,3	665,5	649,0	640,6	602,3	601,2	600,4	594,6	**590,5**
DB	0,453	0,450	0,220	0,228	0,210	0,233	0,282	0,130	**0,121**
Dunn	**2,209**	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
AR	0,002	0,005	0,010	0,012	0,027	0,027	0,027	0,030	**0,032**
RI	0,533	0,533	0,535	0,536	0,542	0,542	0,542	0,543	**0,544**
CA	0,629	0,631	0,634	0,636	0,647	0,647	0,647	0,649	**0,650**

Tab. 4.18 Validační indexy, datový set WDBC (*single-link*)

Val. index	2	3	4	5	6	7	8	9	10
CP	579,2	573,8	326,8	321,3	318,0	225,7	198,8	197,2	**189,4**
DB	0,429	0,410	**0,346**	0,588	0,426	0,398	0,460	0,499	0,531
Dunn	**4,165**	2,244	1,751	2,031	0,000	0,000	0,000	0,000	0,000
AR	0,052	0,052	**0,465**	0,464	0,463	0,421	0,386	0,386	0,379
RI	0,552	0,552	**0,737**	0,736	0,736	0,705	0,686	0,686	0,682
CA	0,663	0,663	**0,854**	**0,854**	**0,854**	**0,854**	**0,854**	**0,854**	**0,854**

Tab. 4.19 Validační indexy, datový set WDBC (*complete-link*)

Val. index	2	3	4	5	6	7	8	9	10
CP	579,2	573,8	321,1	318,6	312,4	310,7	267,4	265,3	**263,6**
DB	0,429	0,410	0,410	**0,374**	0,417	0,418	0,430	0,447	0,425
Dunn	**4,165**	2,244	1,689	0,000	0,000	0,000	0,000	0,000	0,000
AR	0,052	0,052	**0,537**	**0,537**	0,536	0,536	0,490	0,490	0,490
RI	0,552	0,552	**0,771**	**0,771**	**0,771**	**0,771**	0,747	0,747	0,747
CA	0,663	0,663	**0,879**	**0,879**	**0,879**	**0,879**	**0,879**	**0,879**	**0,879**

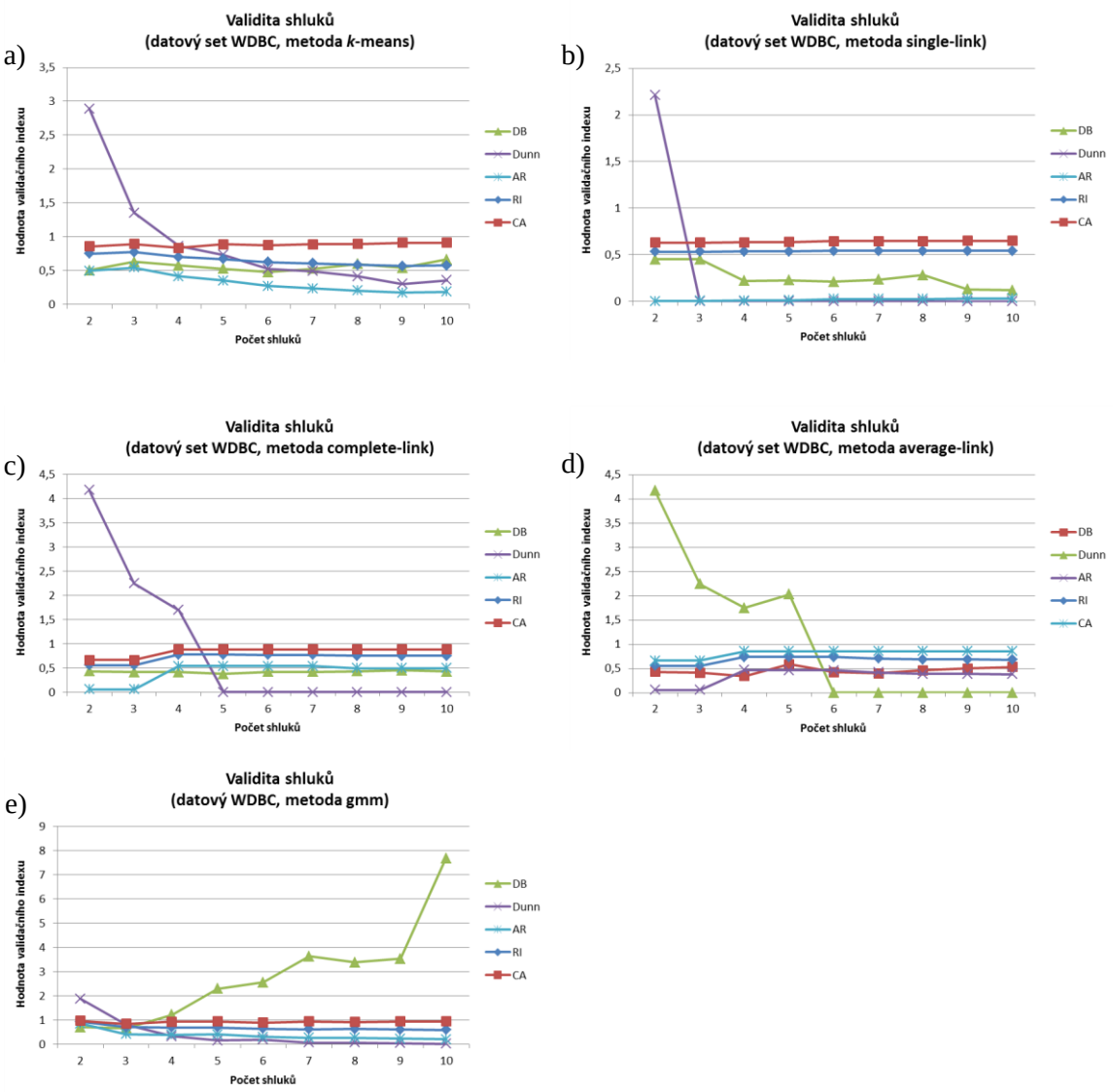
Tab. 4.20 Validační indexy, datový set WDBC (*average-link*)

Val. index	2	3	4	5	6	7	8	9	10

CP	436,3	374,5	411,5	432,0	326,0	382,2	350,4	340,4	**320,3**
DB	0,702	**0,657**	1,213	2,288	2,548	3,626	3,368	3,527	7,673
Dunn	**1,876**	0,808	0,329	0,147	0,187	0,056	0,059	0,041	0,009
AR	**0,844**	0,405	0,383	0,392	0,297	0,263	0,265	0,227	0,204
RI	**0,922**	0,697	0,683	0,687	0,636	0,616	0,618	0,596	0,584
CA	**0,960**	0,837	0,917	0,931	0,882	0,933	0,914	0,937	0,940

Tab. 4.21 Validační indexy, datový set WDBC (*EM*)

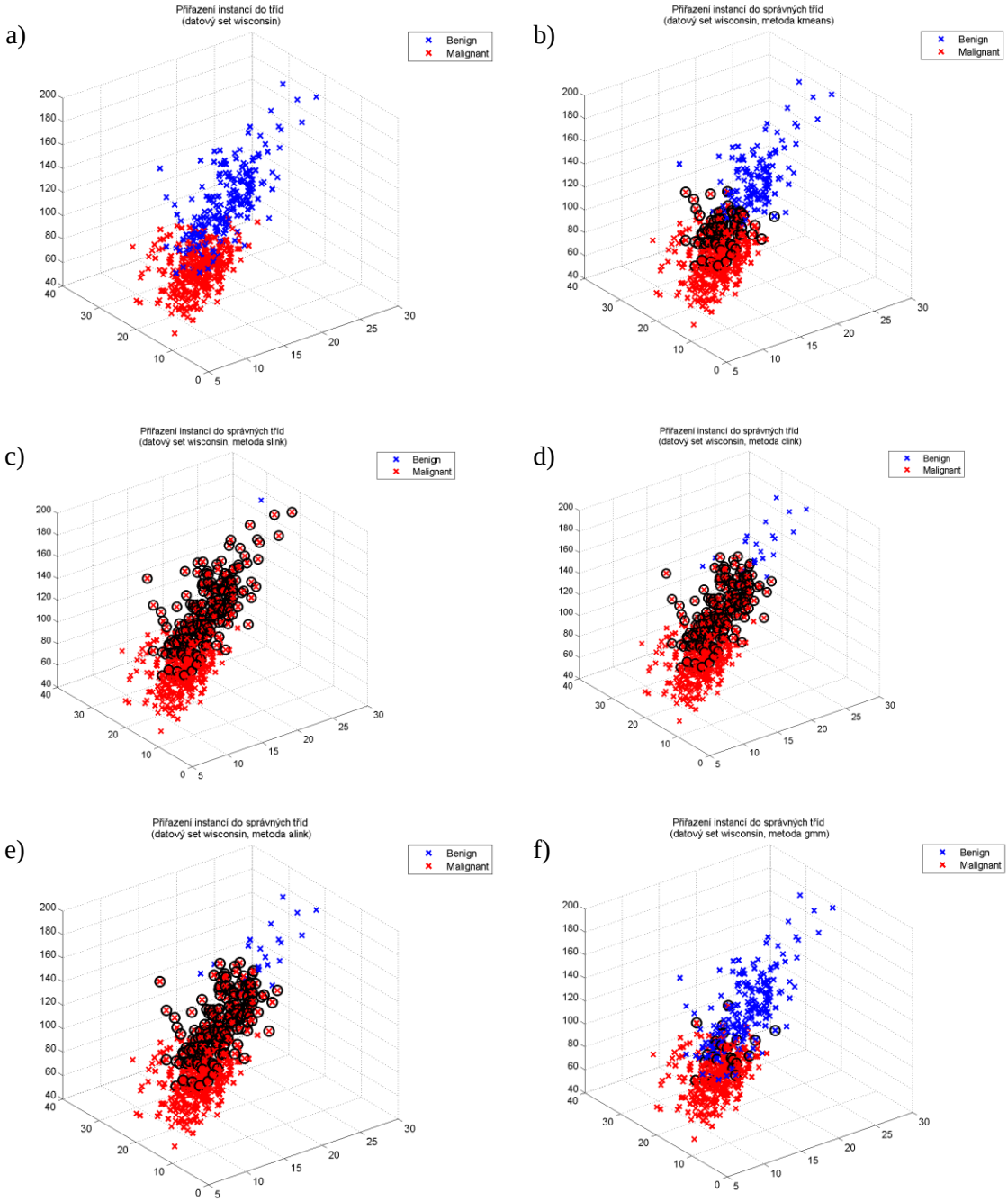
Obrázek 4.8 graficky zobrazuje hodnoty validačních indexů, které jsou uvedeny v tabulkách 4.17 až 4.21. Pro přehlednější prezentaci výsledků není zobrazen index CP. Zvětšený výsek grafů slouží pro přehlednější prezentaci výsledků. Největší počet správných shluků v datovém setu WDBC bylo určeno pomocí Dunnova interního validačního kritéria. Ve všech případech byl určen správný optimální počet shluků datového setu. Externí kritéria AR, RI, CA ohodnotila správné rozložení dat při použití EM algoritmu. Ostatní validační kritéria ani jednou neurčila správný počet shluků předloženého datového setu.



Obr. 4.8 WDBC - validita shluků, a) *k*-means, b) single-link, c) complete-link, d) average-link, e) EM

4.3.3.2 Přiřazení instancí do tříd

Obrázek 4.9 zobrazuje grafické přiřazení instancí do tříd použitého datového setu a přiřazení instancí do tříd po použití jednotlivých shlukovacích algoritmů s vyznačením chybně klasifikovaných bodů černým kroužkem.



Obr. 4.9 WDBC - přiřazení instancí do tříd, a) vstupní data b) *k*-means, c) single-link, d) complete-link, e) average-link, f) EM

Tabulka 4.22 uvádí počty instancí v jednotlivých třídách datového setu WDBC a zařazení instancí do tříd ve výsledcích shlukování danými shlukovacími algoritmy.

Třída	**Benign**	**Malignant**
# instancí	**212**	**357**
k-means	130	1
	82	356
single-link	1	0
	211	357
complete-link	20	0
	192	357
average-link	20	0
	192	357
EM	196	7
	16	350

Tab. 4.22 WDBC - přiřazení instancí do tříd

V tabulce 4.23 jsou celkové součty chybně klasifikovaných přiřazení instancí do tříd a jejich procentuální vyjádření.

	k-means	single-link	complete-link	average-link	EM
Chybně klasifikováno	83	211	192	192	23
Vyjádření v [%]	14,6	37,1	33,7	33,7	4,0

Tab. 4.23 WDBC – chybně přiřazené instance do tříd

Z grafického vyjádření na obrázku 4.9 i číselného vyjádření v tabulkách 4.22 a 4.23 je patrné, že testované shlukovací algoritmy *average-link* a *complete-link* nejsou schopny úspěšně zařadit velké procento instancí datového setu WDBC do správné třídy. Chybovost se u zmíněných algoritmů pohybuje přes 30 %. Kromě *EM* algoritmu všechny ostatní algoritmy správně přiřadily instance do třídy *Malignant*, ale nebyly schopny rozpoznat instance, které se nacházejí v průniku těchto dvou tříd. *EM* algoritmus ale na rozdíl od ostatních algoritmů zařadil téměř všechny správné instance do třídy *Benign*. Nezařadil pouze 23 instancí do správných tříd a chybovost je tak pouhé 4 %. Algoritmus *k*-means který správně přiřadil do třídy *Benign* více než polovinu instancí. Pouze nízké procento (14,6 %) zařadil chybně.

4.3.4 LSVT

4.3.4.1 Stanovení počtu shluků

V tabulkách 4.24 až 4.27 jsou uvedeny hodnoty interních a externích validačních kritérií získaných při evaluaci výsledků shlukování algoritmy *k*-means, *single-link*, *complete-link*, *average-link* při použití datového setu LSVT. *EM* algoritmus nebylo možné pro tento datový set použít, jelikož set obsahuje více atributů než instancí a *EM* algoritmus nebyl schopen z předložených dat utvořit shluky. Zvýrazněny jsou minimální/maximální hodnoty validačních indexů určující stanovený počet shluků.

Val. ind.	2	3	4	5	6	7	8	9	10
CP	9,8E+09	6,96E+09	5,92E+09	4,44E+09	3,67E+09	3,17E+09	2,9E+09	**2,73E+09**	2,84E+09
DB	0,474	0,556	0,517	0,397	0,432	**0,388**	0,394	0,422	0,457
Dunn	**3,080**	1,381	0,942	0,980	0,705	0,455	0,388	0,267	0,170
AR	-0,039	-0,039	-0,032	-0,012	-0,008	-0,002	0,001	**0,009**	0,005
RI	**0,504**	0,481	0,475	0,474	0,468	0,465	0,463	0,466	0,465
CA	0,667	0,667	0,667	0,667	0,667	0,667	0,667	0,667	**0,690**

Tab. 4.24 Validační indexy, datový set LSVT (*k*-means)

Val. ind.	2	3	4	5	6	7	8	9	10
CP	673,3	665,5	649,0	640,6	602,3	601,2	600,4	594,6	**590,5**
DB	0,453	0,450	0,220	0,228	0,210	0,233	0,282	0,130	**0,121**
Dunn	**2,209**	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
AR	0,002	0,005	0,010	0,012	0,027	0,027	0,027	0,030	**0,032**
RI	0,533	0,533	0,535	0,536	0,542	0,542	0,542	0,543	**0,544**
CA	0,629	0,631	0,634	0,636	0,647	0,647	0,647	0,649	**0,650**

Tab. 4.25 Validační indexy, datový set LSVT (*single-link*)

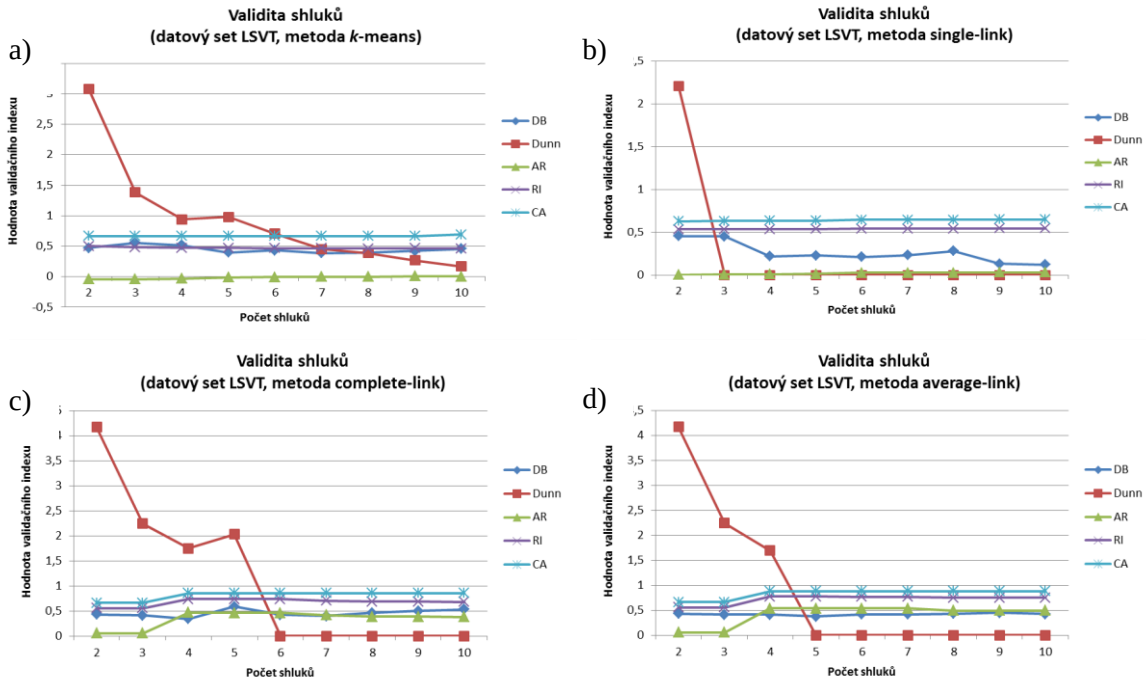
Val. ind.	2	3	4	5	6	7	8	9	10
CP	579,2	573,8	326,8	321,3	318,0	225,7	198,8	197,2	**189,4**
DB	0,429	0,410	**0,346**	0,588	0,426	0,398	0,460	0,499	0,531
Dunn	**4,165**	2,244	1,751	2,031	0,000	0,000	0,000	0,000	0,000
AR	0,052	0,052	**0,465**	0,464	0,463	0,421	0,386	0,386	0,379
RI	0,552	0,552	**0,737**	0,736	0,736	0,705	0,686	0,686	0,682
CA	0,663	0,663	0,854	0,854	**0,854**	**0,854**	**0,854**	**0,854**	**0,854**

Tab. 4.26 Validační indexy, datový set LSVT (*complete-link*)

Val. ind.	2	3	4	5	6	7	8	9	10
CP	579,2	573,8	321,1	318,6	312,4	310,7	267,4	265,3	**263,6**
DB	0,429	0,410	0,410	**0,374**	0,417	0,418	0,430	0,447	0,425
Dunn	**4,165**	2,244	1,689	0,000	0,000	0,000	0,000	0,000	0,000
AR	0,052	0,052	**0,537**	**0,537**	0,536	0,536	0,490	0,490	0,490
RI	0,552	0,552	**0,771**	**0,771**	**0,771**	**0,771**	0,747	0,747	0,747
CA	0,663	0,663	**0,879**	**0,879**	**0,879**	**0,879**	**0,879**	**0,879**	**0,879**

Tab. 4.27 Validační indexy, datový set LSVT (*average-link*)

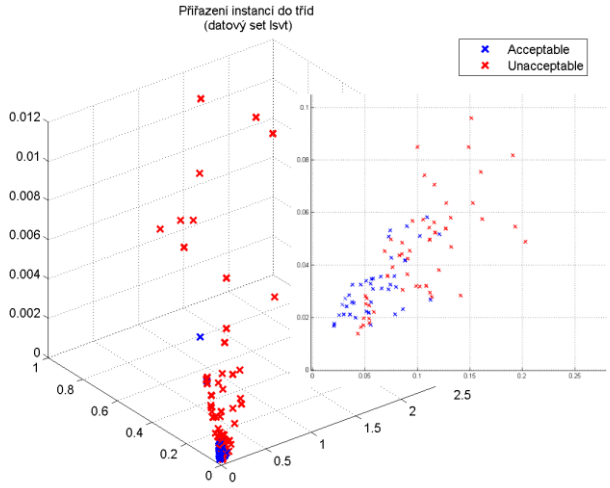
Obrázek 4.10 graficky zobrazuje hodnoty validačních indexů, které jsou uvedeny v tabulkách 4.24 až 4.27. Pro přehlednější prezentaci výsledků není zobrazen index CP. Správný počet shluků byl v každém výsledku shlukování nalezen Dunnovým validačním kritériem. Při shlukování algoritmem *k*-means byl správný počet shluků určen pouze externím kritériem RI.



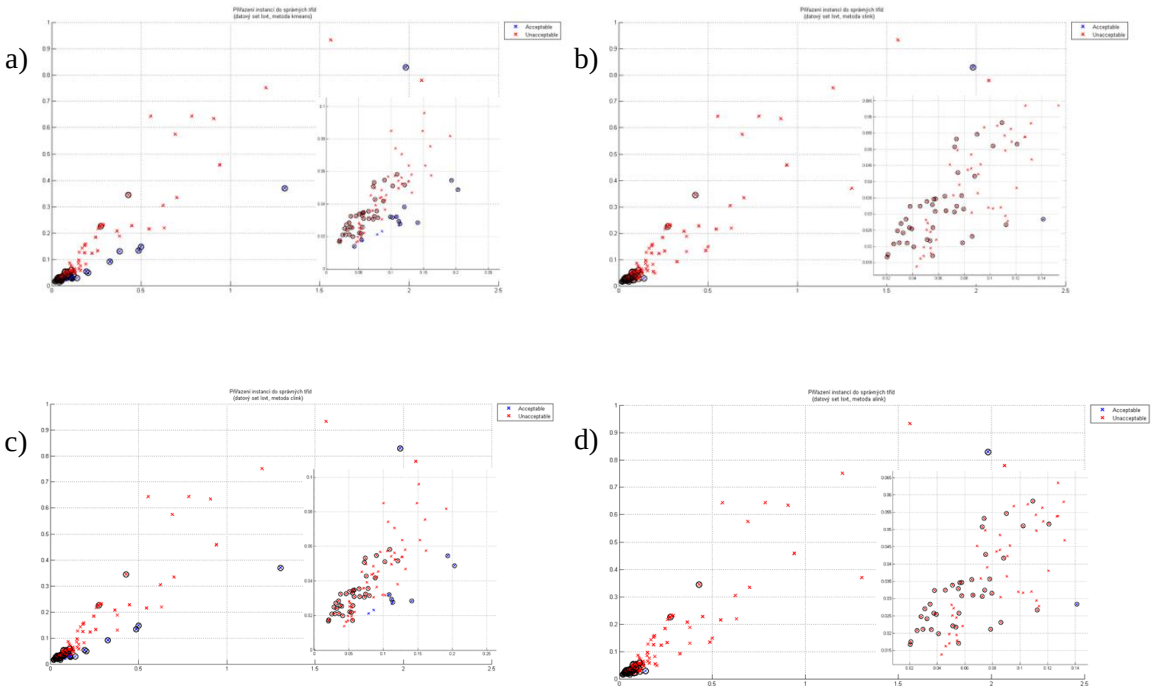
Obr. 4.10 LSVT - validita shluků, a) *k*-means, b) single-link, c) complete-link, d) average-link

4.3.4.2 Přiřazení instancí do tříd

Obrázky 4.11 a 4.12 zobrazují grafické přiřazení instancí do tříd použitého datového setu a přiřazení instancí do tříd po použití jednotlivých shlukovacích algoritmů s vyznačením chybně klasifikovaných bodů. Zvětšený výsek grafů slouží pro přehlednější prezentaci výsledků.



Obr. 4.11 LSVT - přiřazení instancí do tříd



Obr. 4.12 LSVT - přiřazení instancí do správných tříd, a) *k*-means, b) single-link, c) complete-link, d) average-link

Tabulka 4.28 uvádí počty instancí v jednotlivých třídách datového setu LSVT a zařazení instancí do tříd ve výsledcích shlukování danými shlukovacími algoritmy.

Třída	**Acceptable**	**Unacceptable**
# instancí	**42**	**84**
k-means	4	17
	38	67
single-link	0	2
	42	82
complete-link	3	11
	39	73
average-link	0	2
	42	82

Tab. 4.28 LSVT - přiřazení instancí do tříd

V tabulce 4.29 jsou celkové součty chybně klasifikovaných přiřazení instancí do tříd a jejich procentuální vyjádření.

	k-means	single-link	complete-link	average-link
Chybně klasifikováno	55	44	50	44
Vyjádření v [%]	43,7	34,9	39,7	34,9

Tab. 4.29 LSVT – chybně přiřazené instance do tříd

Pro experiment s datovým setem LSVT nebylo možné použít algoritmus *EM*, který nebyl schopen z předložených dat utvořit shluky. Z grafického vyjádření na obrázku 4.12 a číselného vyjádření v tabulkách 4.28 a 4.29 je patrné, že testované algoritmy nebyly schopny úspěšně zařadit správné instance do třídy *Acceptable*. Obě výstupní třídy jsou totiž od sebe málo separovány a velmi těsně se prolínají. Nejlepšího výsledku bylo dosaženo za použití algoritmů *average-link* a *single-link,* které chybně klasifikovaly 34,9 % všech instancí.

4.3.5 Naměřená data

4.3.5.1 Stanovení počtu shluků

V tabulkách 4.30 až 4.34 jsou uvedeny hodnoty interních a externích validačních kritérií získaných při evaluaci výsledků shlukování algoritmy *k*-means, *single-link*,

complete-link, average-link a *EM* algoritmu při použití datového setu získaného měřením anomálií na vibračním přípravku. Zvýrazněny jsou minimální/maximální hodnoty validačních indexů určující stanovený počet shluků.

Val. index	2	3	4	5	6	7
CP	9,6E+05	7,8E+05	6,4E+05	6,0E+05	5,5E+05	5,0E+05
DB	**0,391**	0,595	0,645	0,669	0,690	0,690
Dunn	**4,085**	1,143	1,436	1,699	0,731	0,829
AR	0,234	0,362	0,338	0,310	0,442	**0,477**
RI	0,583	0,752	0,767	0,763	0,847	0,862
CA	0,331	0,481	0,525	0,519	0,657	0,713
Val. index	8	9	10	11	12	13
CP	4,8E+05	4,6E+05	4,7E+05	4,2E+05	4,1E+05	**3,9E+05**
DB	0,687	0,711	0,658	0,783	0,650	0,724
Dunn	0,564	0,588	0,259	0,845	0,556	0,667
AR	0,430	0,409	0,390	0,415	0,377	0,384
RI	0,855	0,858	0,848	0,858	0,850	**0,866**
CA	0,718	0,713	0,718	0,718	0,724	**0,762**

Tab. 4.30 Validační indexy, datový set naměřených dat (*k*-means)

Val. index	2	3	4	5	6	7
CP	2,2E+06	2,1E+06	1,8E+06	1,8E+06	1,8E+06	8,6E+05
DB	0,634	0,633	0,221	0,408	0,263	0,209
Dunn	**1,577**	0,000	0,000	0,000	0,000	0,000
AR	0,000	0,000	0,012	0,011	0,011	**0,242**
RI	0,169	0,177	0,256	0,255	0,262	0,616
CA	0,171	0,177	0,232	0,232	0,238	0,392
Val. index	8	9	10	11	12	13
CP	8,5E+05	8,5E+05	7,9E+05	7,8E+05	7,7E+05	**7,2E+05**
DB	**0,218**	0,265	0,269	0,410	0,404	0,420
Dunn	0,000	0,000	0,000	0,000	0,000	0,000
AR	0,240	0,239	0,225	0,210	0,210	0,193
RI	**0,617**	0,616	0,615	0,615	0,615	0,616
CA	0,398	0,398	0,398	0,398	0,403	**0,420**

Tab. 4.31 Validační indexy, datový set naměřených dat (*single-link*)

Val. index	2	3	4	5	6	7
CP	9,6E+05	8,6E+05	6,8E+05	6,2E+05	5,9E+05	5,6E+05
DB	**0,391**	0,545	0,657	0,703	0,737	0,685
Dunn	**4,085**	2,052	1,261	1,574	1,536	1,536
AR	0,234	0,200	**0,335**	0,311	0,292	0,309
RI	0,583	0,584	0,756	0,756	0,758	0,769
CA	0,331	0,354	0,503	0,508	0,530	0,552
Val. index	8	9	10	11	12	13
CP	5,5E+05	5,0E+05	4,9E+05	4,6E+05	4,5E+05	**4,5E+05**
DB	0,902	0,908	0,853	0,852	0,882	0,938
Dunn	1,121	1,121	1,223	1,223	1,223	1,211
AR	0,302	0,330	0,328	0,294	0,290	0,284
RI	0,768	0,809	0,809	**0,810**	**0,810**	0,809
CA	0,552	0,619	0,619	0,619	**0,624**	**0,624**

Tab. 4.32 Validační indexy, datový set naměřených dat (*complete-link*)

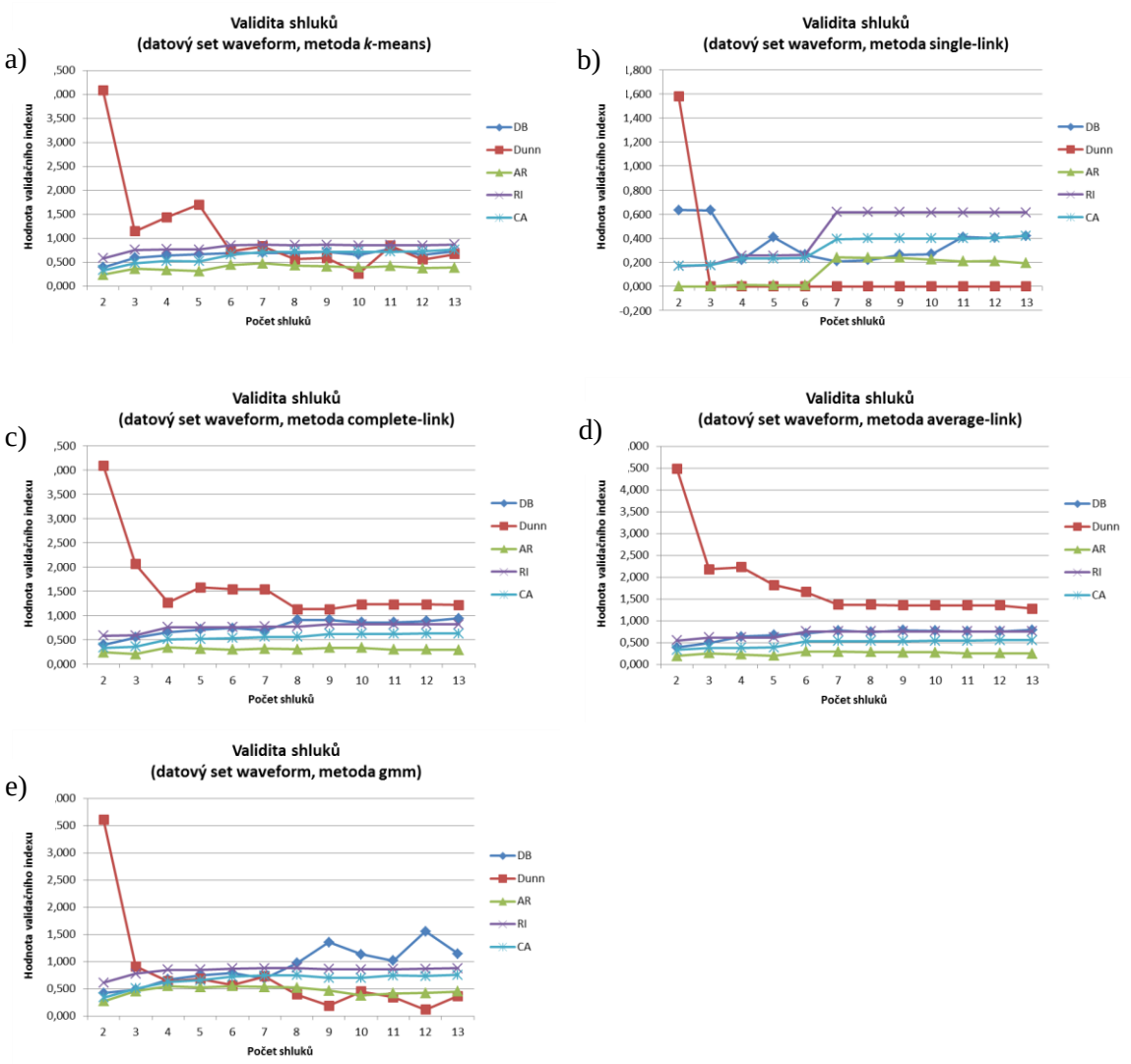
Val. index	2	3	4	5	6	7
CP	1,0E+06	8,8E+05	8,0E+05	7,4E+05	5,9E+05	5,7E+05
DB	**0,384**	0,491	0,636	0,675	0,706	0,773
Dunn	**4,483**	2,180	2,231	1,818	1,659	1,368
AR	0,195	0,254	0,228	0,199	**0,297**	0,288
RI	0,550	0,616	0,616	0,616	0,755	**0,756**
CA	0,331	0,381	0,381	0,387	0,525	0,536
Val. index	8	9	10	11	12	13
CP	5,6E+05	5,5E+05	5,4E+05	5,2E+05	5,2E+05	**5,1E+05**
DB	0,744	0,780	0,771	0,758	0,754	0,794
Dunn	1,368	1,351	1,351	1,351	1,351	1,273
AR	0,283	0,278	0,279	0,258	0,259	0,250
RI	0,754	0,754	0,755	0,753	0,754	0,754
CA	0,536	0,536	0,541	0,541	**0,552**	**0,552**

Tab. 4.33 Validační indexy, datový set naměřených dat (*average-link*)

Val. index	2	3	4	5	6	7
CP	9,9E+05	8,3E+05	7,3E+05	6,9E+05	6,1E+05	5,3E+05
DB	**0,420**	0,482	0,670	0,743	0,791	0,687
Dunn	**3,602**	0,909	0,637	0,684	0,567	0,727
AR	0,273	0,455	0,548	0,525	**0,550**	0,533
RI	0,611	0,779	0,847	0,849	0,872	0,875
CA	0,337	0,503	0,624	0,652	0,724	0,746
Val. index	8	9	10	11	12	13
CP	5,3E+05	5,2E+05	5,1E+05	5,1E+05	5,2E+05	**4,9E+05**
DB	0,970	1,356	1,133	1,016	1,556	1,142
Dunn	0,394	0,189	0,448	0,342	0,118	0,364
AR	0,523	0,463	0,379	0,417	0,419	0,449
RI	**0,881**	0,861	0,853	0,862	0,866	0,875
CA	0,751	0,702	0,702	0,746	0,735	**0,757**

Tab. 4.34 Validační indexy, datový set naměřených dat (*EM*)

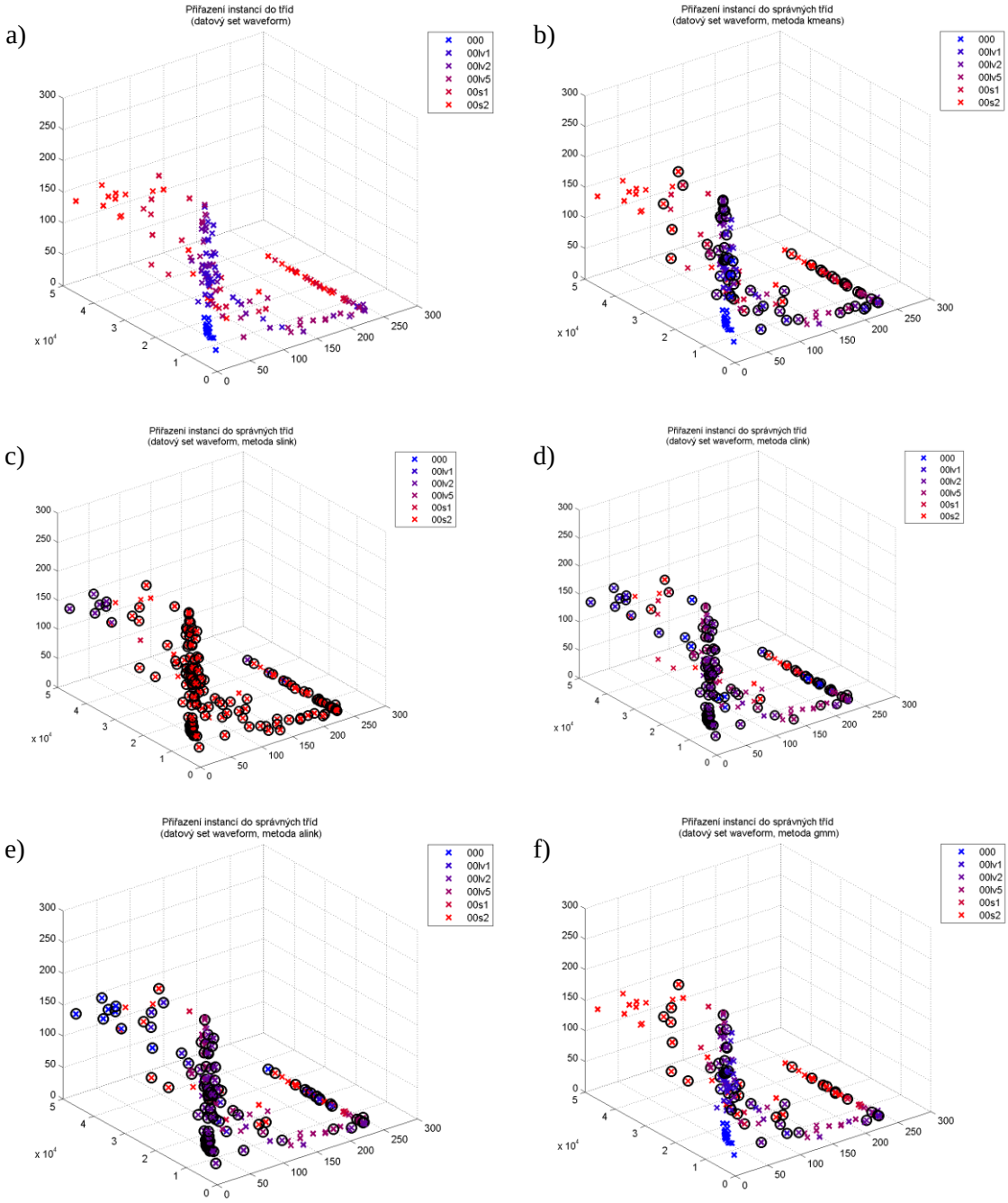
Obrázek 4.13 graficky zobrazuje hodnoty validačních indexů, které jsou uvedeny v tabulkách 4.30 až 4.34. Pro přehlednější prezentaci výsledků není zobrazen index CP. Správný počet shluků datového setu se podařilo určit pouze ve dvou případech indexem AR ze shlukování algoritmy *single-link* a *EM*. V ostatních případech se nepodařilo určit správný počet shluků.



Obr. 4.13 Naměřená data - validita shluků, a) *k*-means, b) single-link, c) complete-link, d) average-link, e) EM

4.3.5.2 Přiřazení instancí do tříd

Obrázek 4.14 zobrazuje grafické přiřazení instancí do tříd použitého datového setu a přiřazení instancí do tříd po použití jednotlivých shlukovacích algoritmů s vyznačením chybně klasifikovaných bodů.



Obr. 4.14 Naměřená data - přiřazení instancí do tříd, a) vstupní data b) *k*-means, c) single-link, d) complete-link, e) average-link, f) EM

Tabulka 4.35 uvádí počty instancí v jednotlivých třídách datového setu naměřených dat a zařazení instancí do tříd ve výsledcích shlukování danými shlukovacími algoritmy.

Třída	**000**	**00lv1**	**00lv2**	**00lv5**	**00s1**	**00s2**
# instancí	**30**	**30**	**30**	**30**	**31**	**30**
k-means	30	2	0	1	0	0
	0	18	10	2	0	0
	0	10	18	11	0	0
	0	0	2	16	4	0
	0	0	0	0	17	10
	0	0	0	0	10	20
single-link	0	0	0	0	1	9
	0	0	0	0	9	10
	30	24	22	5	0	0
	0	6	8	25	0	0
	0	0	0	0	9	0
	0	0	0	0	12	11
complete-link	0	0	0	0	9	5
	0	0	0	0	1	9
	30	24	18	3	0	0
	0	6	12	27	4	0
	0	0	0	0	10	5
	0	0	0	0	7	11
average-link	0	0	0	0	1	9
	0	0	0	0	9	10
	30	24	22	5	0	0
	0	6	8	25	0	0
	0	0	0	0	9	0
	0	0	0	0	12	11
gmm	30	0	0	0	0	0
	0	20	8	1	0	0
	0	10	21	7	0	0
	0	0	1	22	0	0
	0	0	0	0	11	3
	0	0	0	0	20	27

Tab. 4.35 Naměřená data - přiřazení instancí do tříd

V tabulce 4.36 jsou celkové součty chybně klasifikovaných přiřazení instancí do tříd a jejich procentuální vyjádření.

	k-means	single-link	complete-link	average-link	EM
Chybně klasifikováno	62	160	115	114	50
Vyjádření v [%]	34,3	88,4	63,5	63,0	27,6

Tab. 4.36 Naměřená data – chybně přiřazené instance do tříd

Z grafického vyjádření na obrázku 4.14 i číselného vyjádření v tabulkách 4.35 a 4.36 je patrné, že nejvíce instancí byly schopny úspěšně zařadit do tříd shlukovací algoritmy *k*-means (65,7 %) a *EM* (72,4 %). Linkové algoritmy complete-link a average-link nezařadily správně přes 60 % instancí a algoritmus single-link nezařadil téměř 90 % všech instancí. Linkové algoritmy nejsou pro tento typ úloh vhodné.

4.4 Selekce příznaků metodou HFS

Vliv metody selekce příznaků HFS (*Hybrid Feature Selection*) na úspěšnost shlukové analýzy byl experimentálně ověřen na datech získaných z UCI repozitáře a na datech naměřených na vibračním přípravku. Obojí jsou popsána v kapitole 4.2 Metoda selekce příznaků HFS je podrobně představena v kapitole 2.5.1. Pro každý datový set byla na základě naměřených výsledků určena prahová hodnota pro odstranění redundantních příznaků tak, aby výsledná množina příznaků obsahovala takové informace, které jsou nejvíce užitečné pro zařazení instancí do správných výstupních tříd.

Význam (přínos) příznaků pro shlukování datového setu je ohodnocen informačním kritériem symetrické nejistoty SU (*Symmetrical Uncertainity*). Použití metody SU má několik výhod: je ze své podstaty symetrická, tzn. redukuje se počet porovnávaných příznaků, není ovlivněna mnohoznačnými hodnotami příznaků a hodnoty SU jsou normalizovány. Podrobný popis definice kritéria SU je uveden v kapitole 2.5.1. Na základě hodnoty SU je určen *rank* příznaku. Čím je *rank* menší, tím je příznak významnější. Nejvýznamnější příznak bude mít tudíž *rank* = 1. Pre-finální množina příznaků udává pořadí všech příznaků podle jejich významnosti. Významnost se snižuje zleva doprava. Výstupem funkce `hfs` je finální množina příznaků, která je získána odstraněním příznaků s větší absolutní hodnotou korelace, než je zadaná prahová hodnota.

Jak je popsáno v kapitole 4.1, implementace metody pro selekci příznaků HFS v prostředí Matlab je volána funkcí `hfs`. Funkce je volána s variabilními vstupními parametry určujícími vstupní data, počet tříd, počet iterací metody, použité informační kritérium a zvolenou prahovou hodnotu. Výstupními hodnotami jsou: pořadí příznaků

podle jejich významnosti v datovém setu, průměrná hodnota příznaku a standardní odchylka.

Příklad volání funkce:

```
> [r m v] = hfs(data, 3, 300, {'SU'}, 0.695)
```

Voláním funkce `hfs` s výše uvedenými parametry budou příznaky z proměnné *data* klasifikovány do tří tříd a ohodnoceny informačním kritériem SU. Příznaky s absolutní hodnotou korelace větší než 0.695 budou označeny za redundantní a odstraněny.

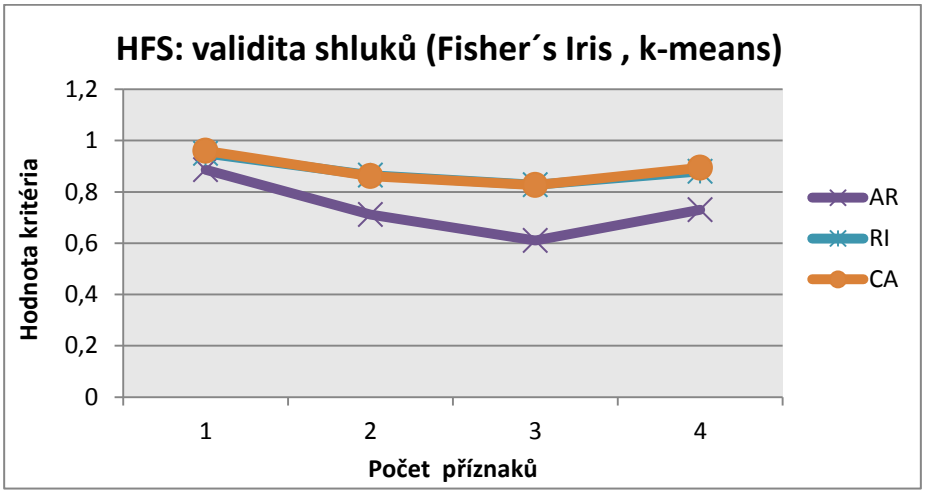
4.4.1 Fisher´s Iris

V tabulce 4.37 jsou uvedeny pozice příznaků v datovém setu Fisher´s Iris podle jejich významnosti. Dále tabulka obsahuje vybranou množinu příznaků získanou odstraněním redundantních příznaků na základě absolutní hodnoty korelace mezi dvojicemi příznaků. Významnost příznaků se snižuje zleva doprava.

Iterativním snižováním prahu pro odstranění redundantních příznaků metodou HFS byla získána redukovaná data, která byla použita pro shlukovou analýzu metodou *k*-means a kvalita výsledných shluků byla ohodnocena externími validačními kritérii. Níže uvedený graf 4.1 zobrazuje výslednou validitu shluků, přičemž uvedené hodnoty jsou zprůměrovanými hodnotami validačních kritérií z deseti iterací shlukovacího algoritmu *k*-means na redukovaných datech. Rozborem grafických výsledků bylo zjištěno, že pro shlukovou analýzu je optimální použit jeden vybraný příznak. Při použití více příznaků se zhoršuje kvalita výsledných shluků a tím se zvyšuje i chybovost přiřazení instancí do správných tříd.

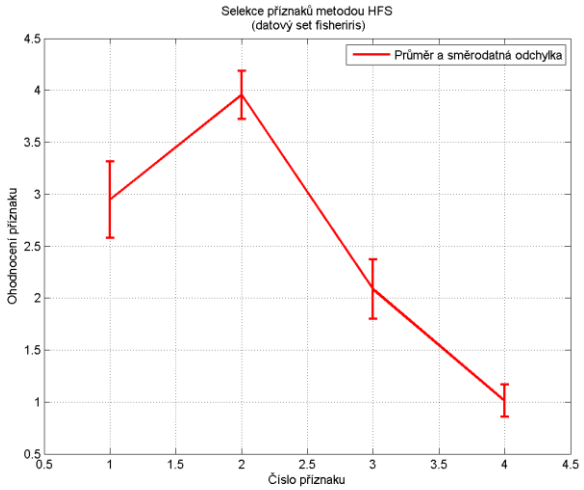
Datový set	Významnost příznaků	Vybrané příznaky	Práh
Fisher´s Iris	{4, 3, 1, 2}	{4}	0,3

Tab. 4.37 Fisher´s Iris - výběr příznaků metodou HFS



Graf 4.1 HFS: validita shluků (Fisher´s Iris, *k*-means)

Obrázek 4.15 ilustruje průměrné ohodnocení (*rank*) příznaků v datovém setu a jejich směrodatnou odchylku.



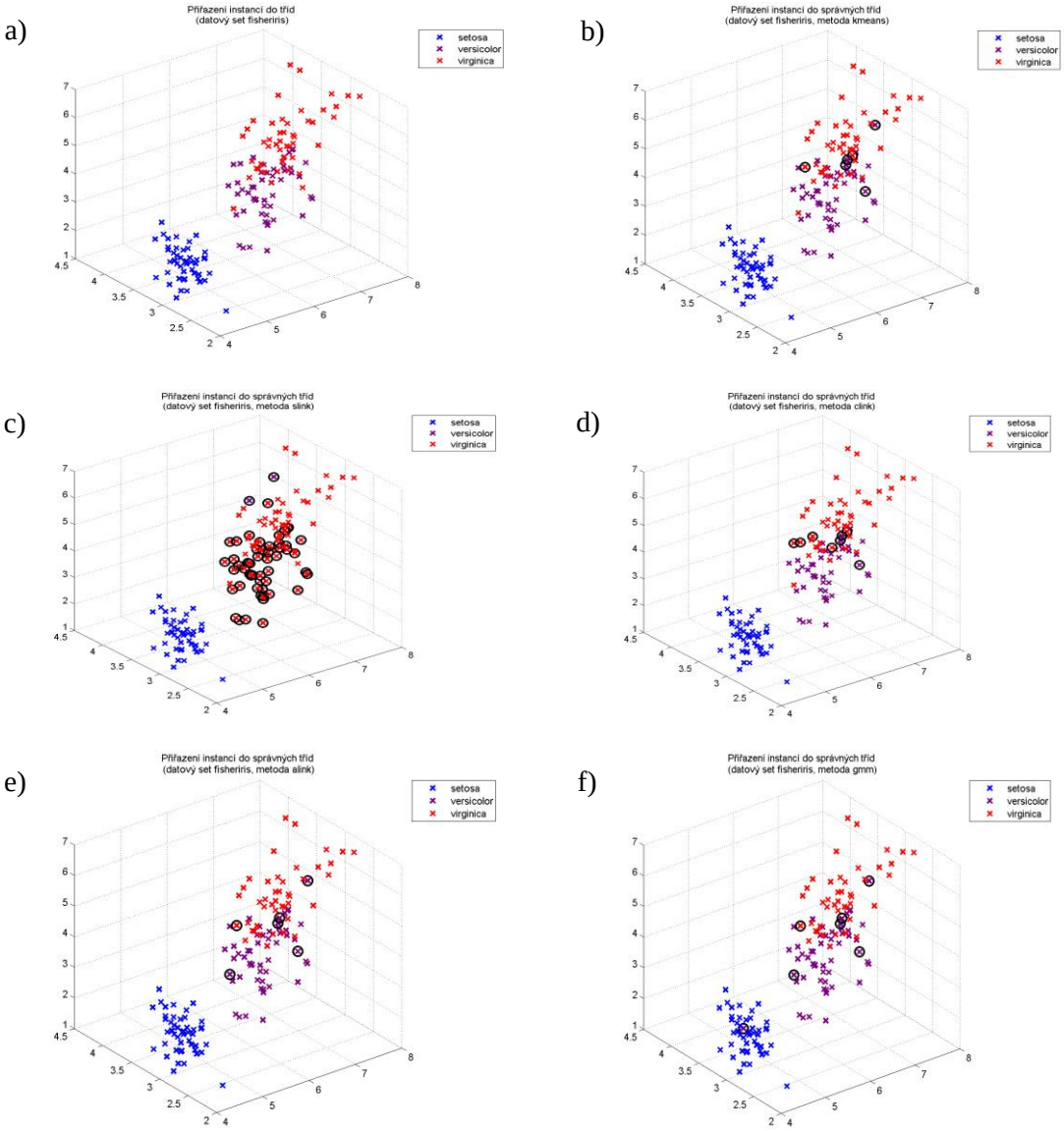
Obr. 4.15 HFS: Fisher´s Iris - průměr a směrodatná odchylka příznaků

	k-means		single-link		complete-link		average-link		EM	
	-	HFS	-	HFS	-	HFS	-	HFS	-	HFS
Chybně klasifikováno	16	6	52	53	24	8	14	6	5	7
[%]	10,6	4,0	34,6	35,3	16,0	5,3	9,3	4,0	3,3	4,6

Tab. 4.38 HFS: Fisher´s Iris – chybně přiřazené instance do tříd

V tabulce 4.38 jsou uvedeny celkové součty chybných přiřazení instancí do tříd a jejich procentuální vyjádření bez selekce příznaků a s použitím metody HFS se zvoleným prahem 0,3 a vybraným jedním příznakem ze čtyř (viz tabulka 4.37).

Obrázek 4.16 zobrazuje grafické přiřazení instancí do tříd použitého datového setu a přiřazení instancí do tříd pomocí shlukovacích algoritmů po použití metody selekce příznaků HSF. Kroužkem jsou vyznačeny chybně klasifikované body.



Obr. 4.16 HFS: Fisher´s Iris - přiřazení instancí do tříd, a) vstupní data b) *k*-means, c) single-link, d) complete-link, e) average-link, f) EM

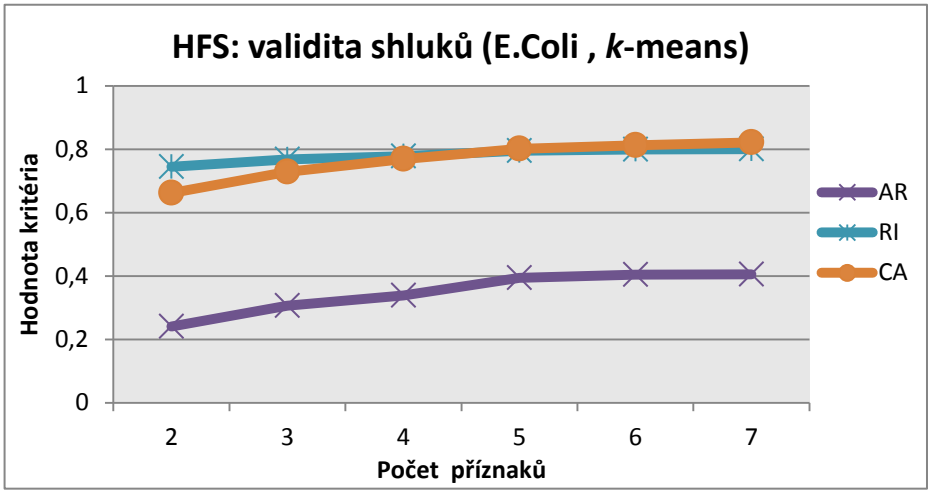
4.4.2 E.coli

V tabulce 4.39 jsou uvedeny pozice příznaků v datovém setu E.coli podle jejich významnosti. Dále obsahuje vybranou množinu příznaků získanou odstraněním redundantních příznaků na základě absolutní hodnoty korelace mezi dvojicemi příznaků. Významnost příznaků se snižuje zleva doprava.

Iterativním snižováním prahu pro odstranění redundantních příznaků metodou HFS byla získána redukovaná data, která byla použita pro shlukovou analýzu metodou *k*-means a kvalita výsledných shluků byla ohodnocena externími validačními kritérii. Níže uvedený graf 4.2 zobrazuje výslednou validitu shluků, přičemž uvedené hodnoty jsou zprůměrovanými hodnotami validačních kritérií z deseti iterací shlukovacího algoritmu *k*-means na redukovaných datech. Rozborem grafických výsledků bylo zjištěno, že pro shlukovou analýzu je optimální použit šest vybraných příznaků. Při použití méně příznaků se zhoršuje kvalita výsledných shluků a tím se zvyšuje i chybovost přiřazení instancí do správných tříd.

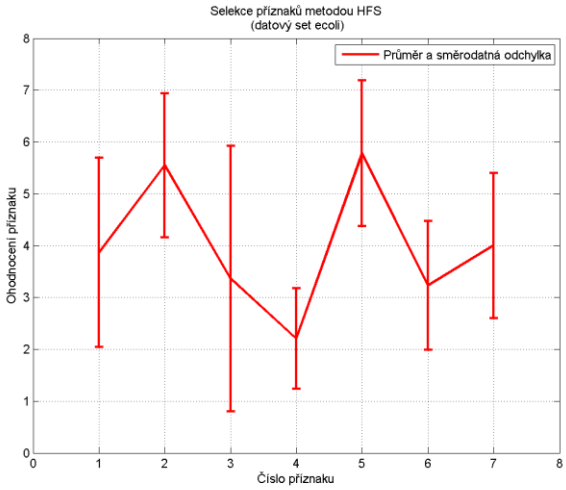
Datový set	Významnost příznaků	Vybrané příznaky	Práh
E.coli	{4, 6, 3, 1, 7, 2, 5}	{4, 6, 3, 1, 2, 5}	0,8

Tab. 4.39 E.coli - výběr příznaků metodou HFS



Graf 4.2 HFS: validita shluků (E.coli, *k*-means)

Obrázek 4.17 ilustruje průměrné ohodnocení (*rank*) příznaků v datovém setu a jejich směrodatnou odchylku.



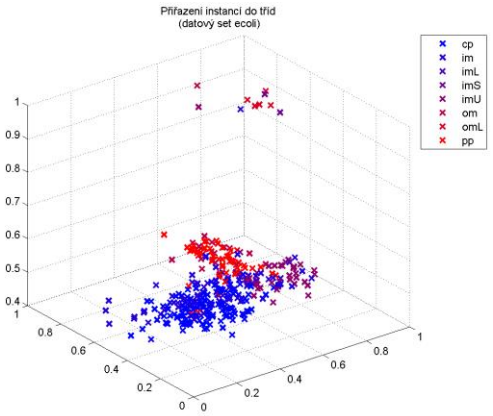
Obr. 4.17 HFS: E.coli - průměr a směrodatná odchylka příznaků

V tabulce 4.40 jsou uvedeny celkové součty chybných přiřazení instancí do tříd a jejich procentuální vyjádření bez selekce příznaků a s použitím metody HFS se zvoleným prahem 0,8 a vybranými šesti příznaky ze sedmi (viz tabulka 4.39).

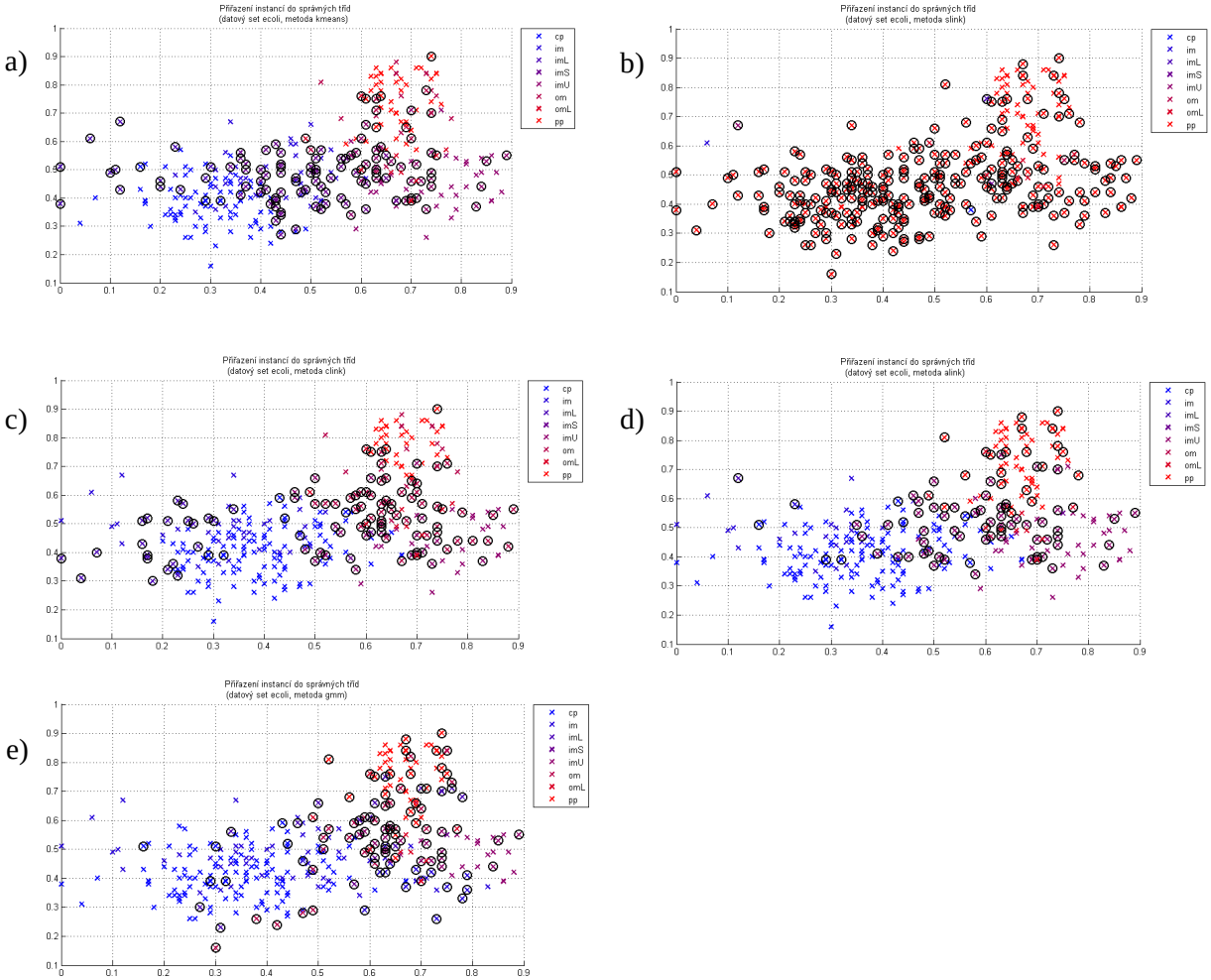
	k-means		single-link		complete-link		average-link		EM	
	-	HFS	-	HFS	-	HFS	-	HFS	-	HFS
Chybně klasifikováno	118	121	278	278	129	108	100	92	122	107
[%]	35,1	36,0	82,7	82,7	38,4	32,1	29,8	27,4	36,3	31,8

Tab. 4.40 HFS: E.coli – chybně přiřazené instance do tříd

Obrázek 4.18 zobrazuje grafické přiřazení instancí do tříd použitého datového setu a obrázek 4.19 přiřazení instancí do tříd pomocí shlukovacích algoritmů po použití metody selekce příznaků HSF. Kroužkem jsou vyznačeny chybně klasifikované body.



Obr. 4.18 E.coli – zařazení instancí do tříd



Obr. 4.19 HFS: E.coli - přiřazení instancí do správných tříd, a) *k*-means, b) single-link, c) complete-link, d) average-link, e) EM

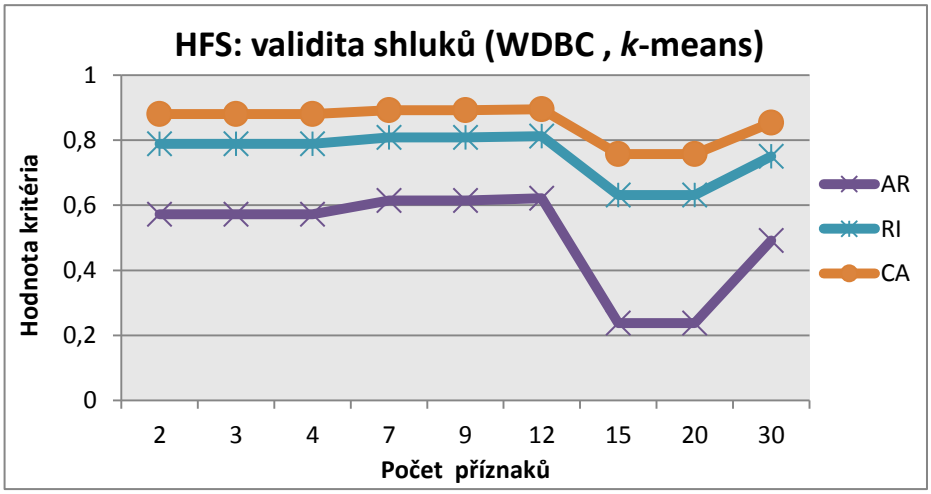
4.4.3 WDBC

V tabulce 4.41 jsou uvedeny pozice příznaků v datovém setu WDBC podle jejich významnosti. Dále obsahuje vybranou množinu příznaků získanou odstraněním redundantních příznaků na základě absolutní hodnoty korelace mezi dvojicemi příznaků. Významnost příznaků se snižuje zleva doprava.

Iterativním snižováním prahu pro odstranění redundantních příznaků metodou HFS byla získána redukovaná data, která byla použita pro shlukovou analýzu metodou *k*-means a kvalita výsledných shluků byla ohodnocena externími validačními kritérii. Níže uvedený graf 4.3 zobrazuje výslednou validitu shluků, přičemž uvedené hodnoty jsou zprůměrovanými hodnotami validačních kritérií z deseti iterací shlukovacího algoritmu *k*-means na redukovaných datech. Rozborem grafických výsledků bylo zjištěno, že pro shlukovou analýzu je optimální použit množinu dvanácti vybraných příznaků. Při použití většího počtu příznaků se výrazně zhoršuje kvalita výsledných shluků a tím se zvyšuje i chybovost přiřazení instancí do správných tříd.

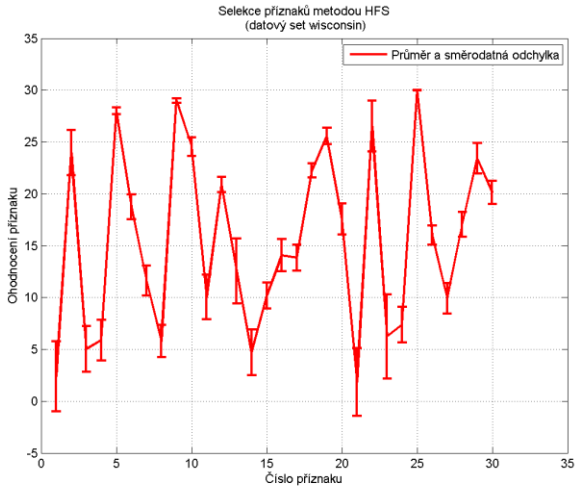
Datový set	Významnost příznaků	Vybrané příznaky	Práh
WDBC	{21, 1, 14, …, 5, 9, 25}	{21, 27, 15, 13, 17, 30, 12, 29, 2, 19, 5, 9}	0,7

Tab. 4.41 WDBC - výběr příznaků metodou HFS



Graf 4.3 HFS: validita shluků (WDBC, *k*-means)

Obrázek 4.20 ilustruje průměrné ohodnocení (*rank*) příznaků v datovém setu a jejich směrodatnou odchylku.



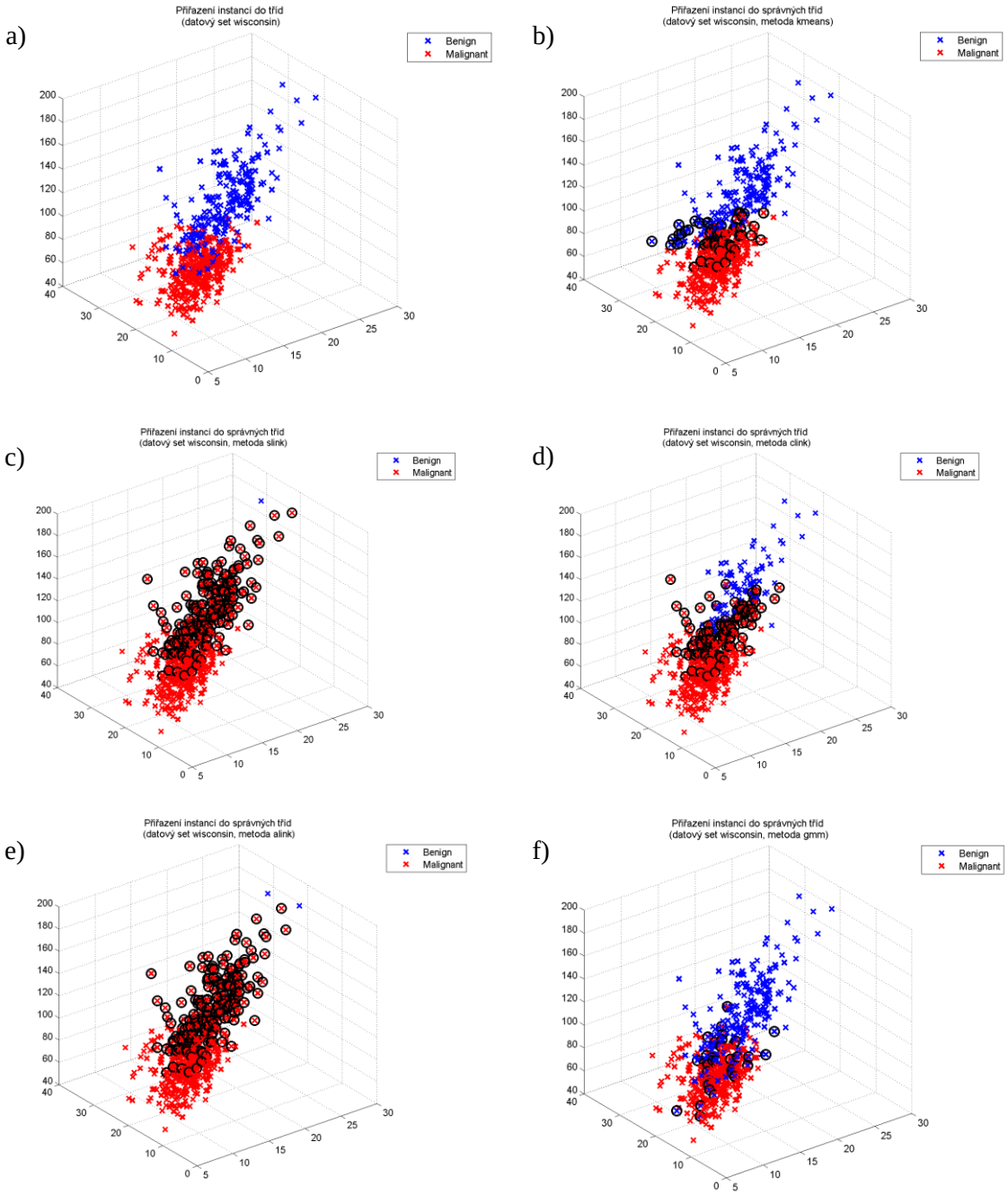
Obr. 4.20 HFS: WDBC - průměr a směrodatná odchylka příznaků

V tabulce 4.42 jsou uvedeny celkové součty chybných přiřazení instancí do tříd a jejich procentuální vyjádření bez selekce příznaků a s použitím metody HFS se zvoleným prahem 0,7 a vybranými dvanácti příznaky z třiceti (viz tabulka 4.41).

	k-means		single-link		complete-link		average-link		EM	
	-	HFS	-	HFS	-	HFS	-	HFS	-	HFS
Chybně klasifikováno	83	59	211	211	192	123	192	210	23	33
[%]	14,6	10,3	37,1	37,1	33,7	21,6	33,7	36,9	4,0	5,8

Tab. 4.42 HFS: WDBC – chybně přiřazené instance do tříd

Obrázek 4.21 zobrazuje grafické přiřazení instancí do tříd použitého datového setu a přiřazení instancí do tříd po použití jednotlivých shlukovacích algoritmů s vyznačením chybně klasifikovaných bodů černým kroužkem.



Obr. 4.21 HFS: WDBC - přiřazení instancí do tříd, a) vstupní data b) *k*-means, c) single-link, d) complete-link, e) average-link, f) EM

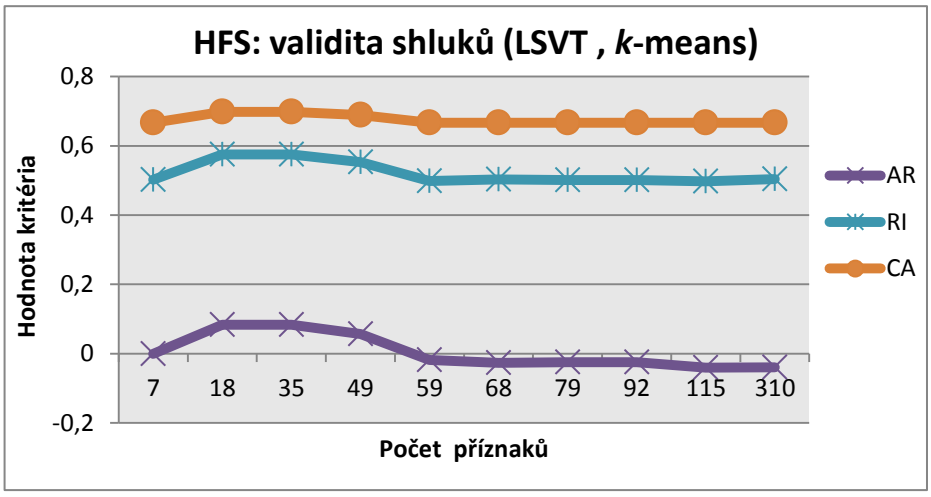
4.4.4 LSVT

V tabulce 4.43 jsou uvedeny pozice příznaků v datovém setu LSVT podle jejich významnosti. Dále tabulka obsahuje vybranou množinu příznaků získanou odstraněním redundantních příznaků na základě absolutní hodnoty korelace mezi dvojicemi příznaků. Významnost příznaků se snižuje zleva doprava.

Iterativním snižováním prahu pro odstranění redundantních příznaků metodou HFS byla získána redukovaná data, která byla použita pro shlukovou analýzu metodou *k*-means a kvalita výsledných shluků byla ohodnocena externími validačními kritérii. Níže uvedený graf 4.4 zobrazuje výslednou validitu shluků, přičemž uvedené hodnoty jsou zprůměrovanými hodnotami validačních kritérií z deseti iterací shlukovacího algoritmu *k*-means na redukovaných datech. Rozborem grafických výsledků bylo zjištěno, že pro shlukovou analýzu je optimální použit osmnáct vybraných příznaků. Při použití více příznaků se mírně zhoršuje kvalita výsledných shluků a tím se zvyšuje i chybovost přiřazení instancí do správných tříd.

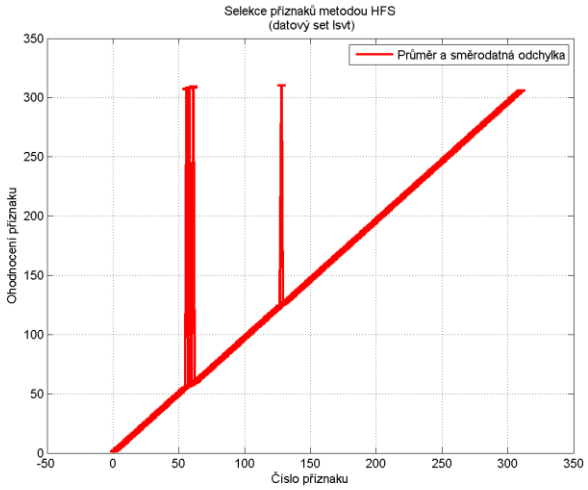
Datový set	Významnost příznaků	Vybrané příznaky	Práh
WDBC	{1, 2, 3, … , 58, 61, 128}	{1, 4, 34, 48, 64, 71, 90, 96, 97, 100, 103, 104, 106, 108, 109, 111, 113, 235}	0,2

Tab. 4.43 LSVT - výběr příznaků metodou HFS



Graf 4.4 HFS: validita shluků (LSVT, *k*-means)

Obrázek 4.22 ilustruje průměrné ohodnocení (*rank*) příznaků v datovém setu a jejich směrodatnou odchylku.



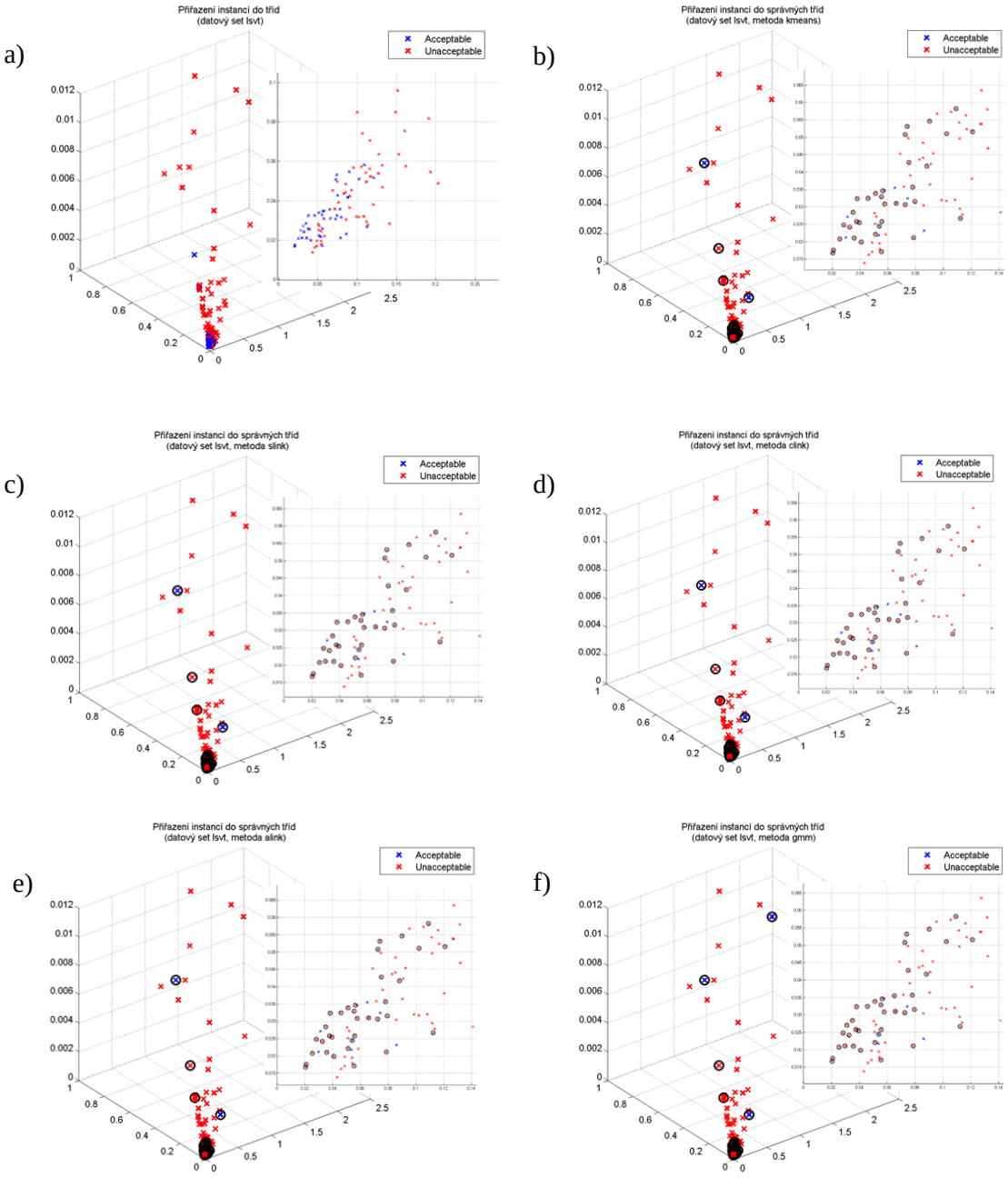
Obr. 4.22 HFS: LSVT - průměr a směrodatná odchylka příznaků

V tabulce 4.44 jsou uvedeny celkové součty chybných přiřazení instancí do tříd a jejich procentuální vyjádření bez selekce příznaků a s použitím metody HFS se zvoleným prahem 0,2 a vybranými osmnácti příznaky z 309 (viz tabulka 4.43).

	k-means		single-link		complete-link		average-link		EM	
	-	HFS	-	HFS	-	HFS	-	HFS	-	HFS
Chybně klasifikováno	55	38	44	39	50	39	44	38	-	42
[%]	43,7	30,2	34,9	31,0	39,7	31,0	34,9	30,2	-	33,3

Tab. 4.44 HFS: LSVT – chybně přiřazené instance do tříd

Obrázek 4.23 zobrazuje grafické přiřazení instancí do tříd použitého datového setu a přiřazení instancí do tříd po použití jednotlivých shlukovacích algoritmů s vyznačením chybně klasifikovaných bodů černým kroužkem. Zvětšený výsek grafů slouží pro přehlednější prezentaci výsledků.



Obr. 4.23 HFS: LSVT - přiřazení instancí do tříd, a) vstupní data b) *k*-means, c) single-link, d) complete-link, e) average-link, f) EM

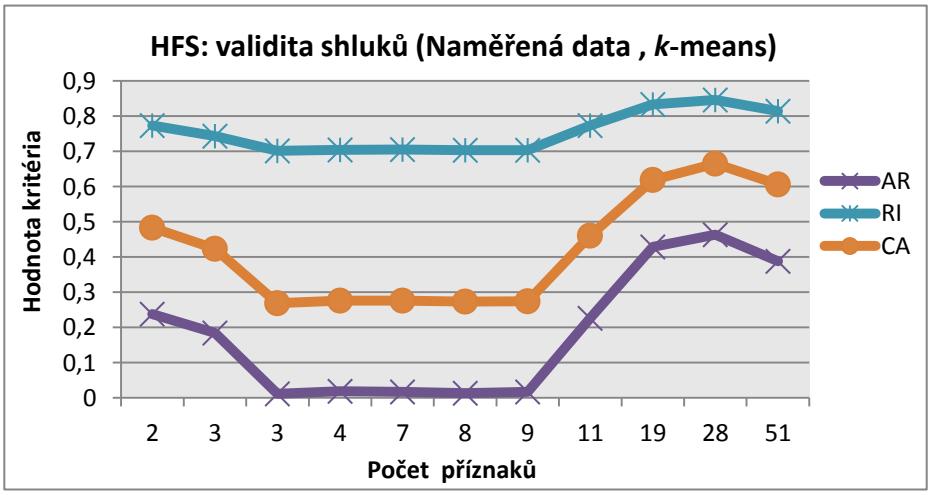
4.4.5 Naměřená data

V tabulce 4.45 jsou uvedeny pozice příznaků v datovém setu naměřených dat podle jejich významnosti. Dále tabulka obsahuje vybranou množinu příznaků získanou odstraněním redundantních příznaků na základě absolutní hodnoty korelace mezi dvojicemi příznaků. Významnost příznaků se snižuje zleva doprava.

Iterativním snižováním prahu pro odstranění redundantních příznaků metodou HFS byla získána redukovaná data, která byla použita pro shlukovou analýzu metodou *k*-means a kvalita výsledných shluků byla ohodnocena externími validačními kritérii. Níže uvedený graf 4.5 zobrazuje výslednou validitu shluků, přičemž uvedené hodnoty jsou zprůměrovanými hodnotami validačních kritérií z deseti iterací shlukovacího algoritmu *k*-means na redukovaných datech. Rozborem grafických výsledků bylo zjištěno, že pro shlukovou analýzu je optimální použit 28 vybraných příznaků z 51. Při nižším počtu použitých příznaků se velmi výrazně zhoršuje kvalita výsledných shluků a tím se zvyšuje i chybovost přiřazení instancí do správných tříd.

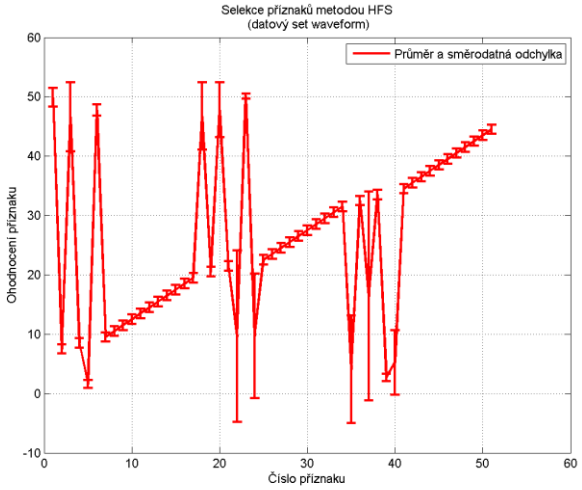
Datový set	Významnost příznaků	Vybrané příznaky	Práh
Naměřená data	{5, 35, 39, … , 6, 1, 23}	{5, 39, 35, 2, 4, 22, 7, 24, 8, 37, 16, 19, 21, 25, 33, 34, 36, 41, 42, 43, 45, 48, 18, 3, 6, 20, 1, 23}	0,96

Tab. 4.45 Naměřená data - výběr příznaků metodou HFS



Graf 4.5 HFS: validita shluků (naměřená data, *k*-means)

Obrázek 4.24 ilustruje průměrné ohodnocení (*rank*) příznaků v datovém setu a jejich směrodatnou odchylku.



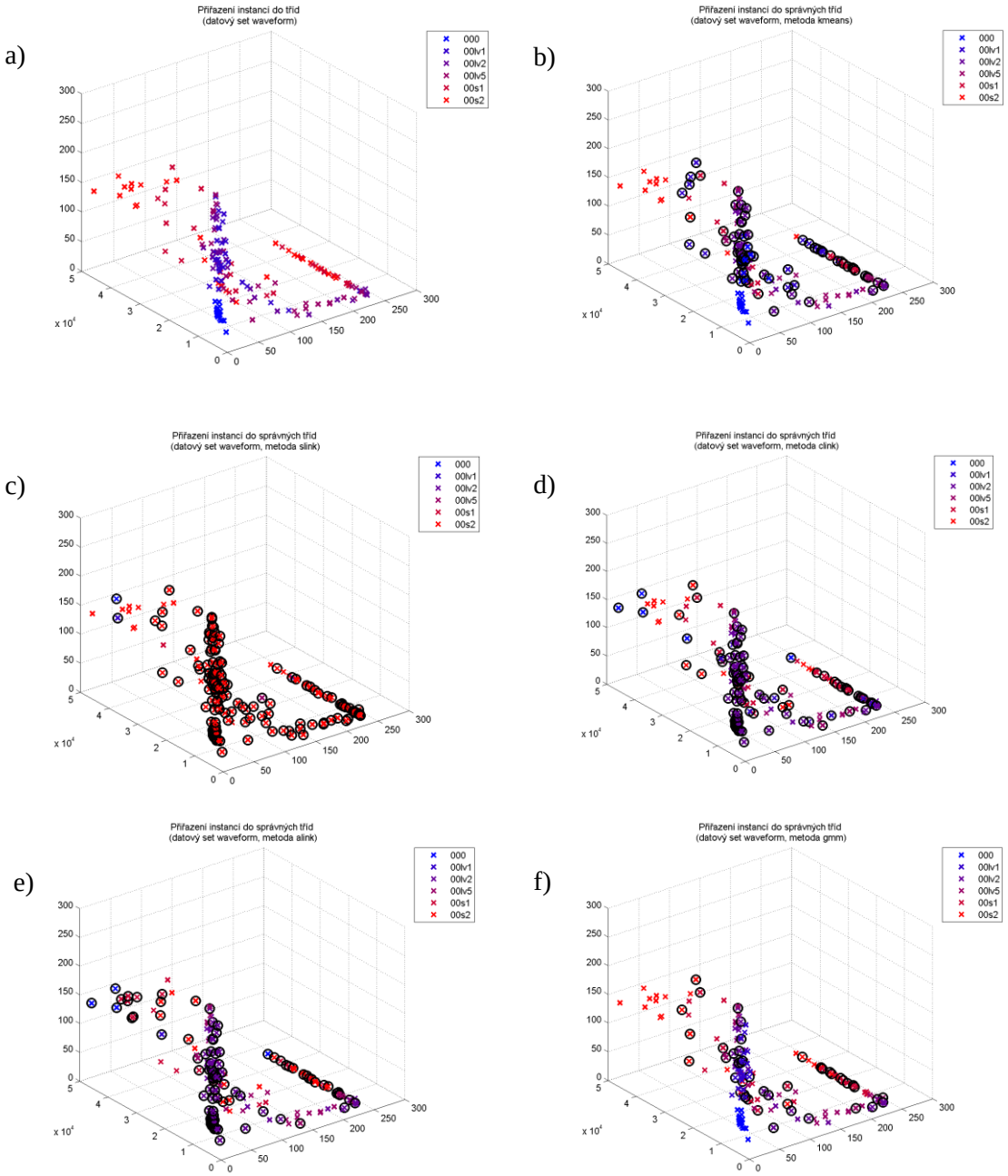
Obr. 4.24 HFS: Naměřená data - průměr a směrodatná odchylka příznaků

V tabulce 4.46 jsou uvedeny celkové součty chybných přiřazení instancí do tříd a jejich procentuální vyjádření bez selekce příznaků a s použitím metody HFS se zvoleným prahem 0,96 a vybranými 28 příznaky uvedenými v tabulce 4.45.

	k-means		single-link		complete-link		average-link		EM	
	-	HFS	-	HFS	-	HFS	-	HFS	-	HFS
Chybně klasifikováno	62	55	160	154	115	110	114	106	50	38
[%]	34,3	30,4	88,4	85,1	63,5	60,8	63,0	58,6	27,6	21,0

Tab. 4.46 HFS: Naměřená data – chybně přiřazené instance do tříd

Obrázek 4.25 zobrazuje grafické přiřazení instancí do tříd použitého datového setu a přiřazení instancí do tříd po použití jednotlivých shlukovacích algoritmů s vyznačením chybně klasifikovaných bodů černým kroužkem.



Obr. 4.25 HFS: Naměřená data - přiřazení instancí do tříd, a) vstupní data b) *k*-means, c) single-link, d) complete-link, e) average-link, f) EM

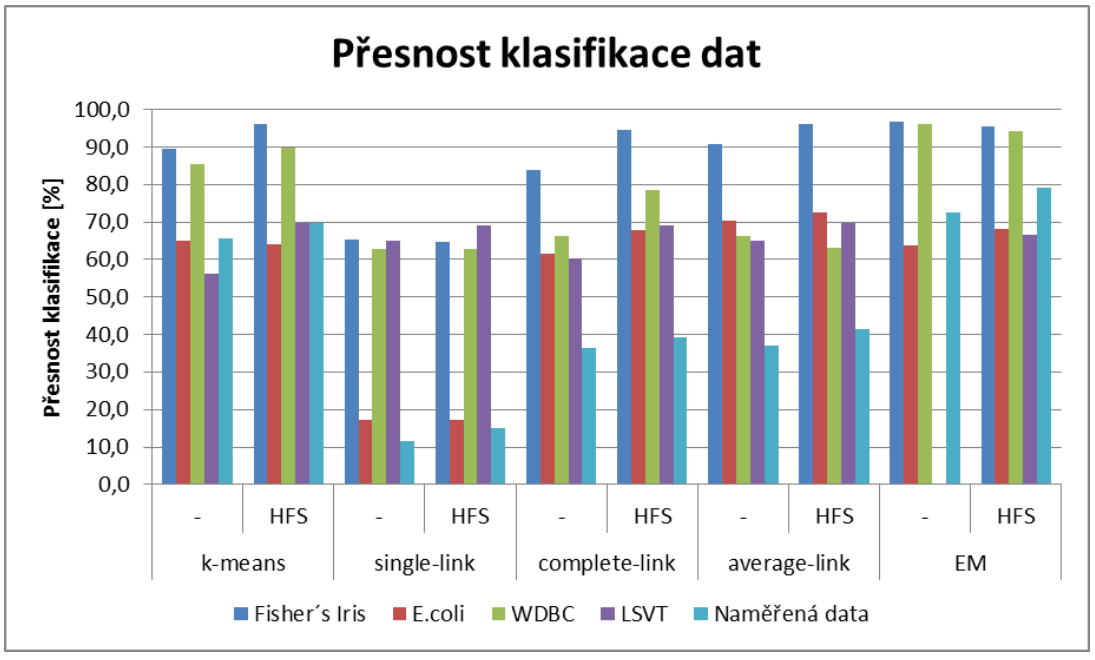
4.5 Přínos metody HFS

Datové sety s tisíci nebo miliony proměnných (příznaků) jsou v dnešní době docela obvyklé. Ne všechna data nesou v sobě využitelné informace, taková data jsou pro nás nadbytečná a je vhodné je odstranit, selektovat. Metoda HFS (*Hybrid Feature Selection*) provádí selekci příznaků na základě jejich významu pro datový set. Vybírá pouze ty příznaky, které urychlí a zjednoduší následné operace s daty.

Úspěšnost metody HFS nejlépe charakterizuje souhrnná přehledná tabulka 4.47 a její grafické znázornění v grafu 4.6.

	k-means		single-link		complete-link		average-link		EM	
	-	HFS	-	HFS	-	HFS	-	HFS	-	HFS
Fisher´s Iris	89,4	96,0	65,4	64,7	84,0	94,7	90,7	96,0	96,7	95,4
E.coli	64,9	64,0	17,3	17,3	61,6	67,9	70,2	72,6	63,7	68,2
WDBC	85,4	89,7	62,9	62,9	66,3	78,4	66,3	63,1	96,0	94,2
LSVT	56,3	69,8	65,1	69,0	60,3	69,0	65,1	69,8	0,0	66,7
Naměřená data	65,7	69,6	11,6	14,9	36,5	39,2	37,0	41,4	72,4	79,0

Tab. 4.47 Přesnost klasifikace dat



Graf 4.6 Přesnost klasifikace dat

Data v datovém setu Fisher´s Iris jsou klasifikována do tří tříd, z nichž jedna třída je od druhých výrazně separována. S výjimkou algoritmu single-link použité algoritmy přiřadily instance do správných tříd s poměrně vysokou přesností. Nejlepšího výsledku dosáhl EM algoritmus (96,7 %). Selekcí příznaků metodou HFS se u algoritmů *k*-means, *complete-link* a *average-link* významně zvýšila přesnost řazení instancí do tříd, a to až o téměř 10 %.

Dalším datovým setem, na kterém je provedeno testování metody HFS, je datový set E.coli. Data jsou klasifikována do osmi tříd, které se nacházejí v těsné blízkosti. Je proto velmi obtížné zařadit data shlukovacími algoritmy do správných tříd. Selekcí příznaků metodou HFS se u algoritmů *complete-link, average-link* a *EM* se zvýšila přesnost řazení instancí do tříd v průměru o 6 %.

Třetím testovaným datovým setem je WDBC. Data jsou roztříděna do dvou tříd, jež jsou od sebe částečně separovány, ale část dat se nachází v jejich průniku. Nejlepších výsledků při klasifikaci dat do správných tříd dosáhly algoritmy *k*-means (85,4 %) a *EM* (96,0 %). Úspěšnost přiřazení dat do tříd bez použití metody HFS byla u algortimu *complete-link* 66,3 %. Selekcí příznaků metodou HFS se u algoritmů *k*-means a *complete-link* zvýšila přesnost řazení instancí do tříd na 89,7 %, respektive 78,4 %. Metoda HFS byla v případě *EM* algoritmu méně úspěšná o 1,8 %.

Čtvrtým testovaným datovým setem je datový set LSVT, v němž existují dvě třídy dat, které se z velké části prolínají. Proto je také obtížná klasifikace dat do správných tříd. Datový set obsahuje více příznaků než instancí. To je důvod, proč *EM* algoritmus nebyl schopen z předložených dat utvořit shluky. Selekcí příznaků metodou HFS byla data redukována, což vedlo k možnosti použít EM algoritmus ke shlukování. U všech testovaných algoritmů došlo po použití metody HFS k nárůstu procentuální úspěšnosti klasifikace dat do tříd. Nejvýrazněji u *EM* algoritmu (66,7 %) a algoritmu *k*-means (13,5 %).

Posledním testovaným datovým setem je soubor naměřených dat na vibračním přípravku. Algoritmus *single-link* dosahuje při klasifikaci instancí do tříd zdaleka nejhorších výsledků, jeho úspěšnost je 11,6 %. Ani po použití metody HFS nedochází k výraznému nárůstu schopnosti klasifikovat data do tříd. Data jsou velmi málo strukturovaná, algoritmus pro ně není vhodný. Nejlepšího výsledku dosáhl algoritmus *EM* se 79 % úspěšnosti. I u dalších algoritmů se zvýšila procentuální klasifikace instancí do tříd, ale jen mírně (o 3 až 7 %).

Metoda HFS pro selekci příznaků prokázala význam redukce datových souborů o redundantní příznaky. U všech testovaných shlukovacích algoritmů na vybraných datových setech se ve většině případů projevilo výrazné zlepšení klasifikace instancí do správných tříd.

5 ZÁVĚR

Diplomová práce na téma Vliv selekce příznaků metodou HFS na shlukovou analýzu je rozdělena do tří logických celků - principy shlukové analýzy, metody shlukové analýzy a experimentální srovnání metod shlukové analýzy.

První část (kap. 2) je teoretickým úvodem do problematiky shlukové analýzy. Definuje pojem shluková analýza, zmiňuje její využití v praktických úlohách v různých vědních oborech, věnuje se matematické formulaci úloh shlukové analýzy, definuje míry podobnosti a vzdálenosti dvou objektů a v závěru zpracovává dvě důležitá témata - stanovení optimálního počtu shluků a selekce příznaků. V podkapitole stanovení optimálního počtu shluků (kap. 2.4) jsou uvedena validační kritéria využívaná k ohodnocení výsledku shlukování, z nichž některá byla vybrána pro experimentální srovnání metod shlukové analýzy. V podkapitole selekce příznaků (kap. 2.5) je zpracován stručný přehled metod selekce příznaků v úlohách bez učitele. Zvláštní pozornost je věnována metodě HFS, která je zde podrobněji teoreticky popsána.

Druhá část (kap. 3) se zaměřuje na přehled metod shlukové analýzy. Definuje pět základních kategorií - metody rozkladu, hierarchické metody, metody založené na hustotě, metody založené na mřížce a metody založené na modelu. Uvádí rovněž výhody a nevýhody každé z uvedených kategorií.

Zásadní částí diplomové práce je experimentální srovnání metod shlukové analýzy (kap. 4). Kapitola obsahuje tři stěžejní podkapitoly - stanovení počtu shluků a zařazení instancí do tříd, selekce příznaků metodou HFS a přínos metody HFS.

V podkapitole stanovení počtu shluků a zařazení instancí do tříd (kap. 4.3) jsou interními a externími validačními kritérii ohodnoceny výsledky shlukování a na jejich základě je stanoven optimální počet shluků pro daný datový set. Z výsledků je patrné, že hlavní nevýhodou použitých validačních kritérií je, že neumí určit správné rozložení shluků, jestliže shluky dat nejsou mezi sebou dostatečně separovány. Testování je provedeno na datech se známou výstupní třídou. To umožnilo porovnat výsledky shlukování a určit úspěšnost zařazení instancí do správných tříd.

V podkapitole selekce příznaků metodou HFS (kap. 4.4) je popsán program, který byl vyvinut v rámci diplomové práce. Srovnává úspěšnost selekce příznaků metodou HFS s výsledky shlukování pěti shlukovacích algoritmů (*k*-means, *single-link, complete-link, average-link* a *EM*). Algoritmus *k*-means byl použit jako základní algoritmus pro určení prahu pro odstranění redundantních dat. V téže podkapitole je uvedeno experimentální srovnání úspěšnosti metody HFS pro selekci příznaků s výsledky shlukování dat bez selekce příznaků. Experiment byl proveden na pěti vybraných datových setech, z nichž jeden vznikl naměřením časových průběhů na vibračním přípravku. Vybrané datové sety reprezentují datové sety s různými charakteristikami, jako jsou počet příznaků a počet instancí.

V podkapitole přínos metody HFS (kap. 4.5) jsou podrobně rozebrány výsledky experimentálního ověření metody HFS pro selekci příznaků. Bylo zjištěno, že metoda HFS ve většině případů výrazně ovlivní výsledek shlukování. Má pozitivní vliv na časovou náročnost a úspěšnost shlukování.

V budoucnu by bylo možné vytvořit grafickou nadstavbu programu a implementovat další shlukovací metody a klasifikační kritéria.

LITERATURA

[1] HAN, Jiawei, Micheline KAMBER a Jian PEI. *Data Mining: Concepts and Techniques*. 3rd ed. Waltham: Morgan Kaufmann, 2012, 703 s. ISBN 978-0-12-381479-1.

[2] TAN, Pang-Ning, Vipin KUMAR a Michael STEINBACH. *Introduction to Data Mining*. 1st ed. Boston: Pearson Addison Wesley, 2006, 769 s. ISBN 0321321367.

[3] RAFSANJANI, Marjan Kuchaki, Zahra Asghari VARZANEH a Nasibeh Emami CHUKANTO. A Survey of Hierarchical Clustering Algorithms. In: *The Journal of Mathematics and Computer Science* [online]. 2012, vol. 5, no. 3, s. 229-240 [cit. 2015-05-18]. Dostupné z: http://www.tjmcs.com

[4] ŘEZÁNKOVÁ, Hana, Dušan HÚSEK a Václav SNÁŠEL. *Shluková analýza dat*. 1. vyd. Praha: Professional Publishing, 2007, 196 s. ISBN 978-80-86946-26-9.

[5] HEBÁK, Petr. *Statistické myšlení a nástroje analýzy dat*. 1. vyd. Praha: Informatorium, 2013, 877 s. ISBN 978-80-7333-105-4.

[6] MELOUN, Milan a Jiří MIITKÝ. Přednosti analýzy shluků ve vícerozměrné statistické analýze. *Zajištění kvality analytických výsledků: sborník přednášek ze semináře 22.-24.3.2003 [na Medlově]*. Český Těšín: 2 Theta, 2004, s. 29-46. ISBN 80-86380-22-X.

[7] HOLČÍK, Jiří. *Analýza a klasifikace dat*. 1. vyd. Brno: Akademické nakladelství CERM, 2012, 111s. ISBN 978-80-7204-793-2.

[8] KELBEL, Jan a David ŠILHÁN. *Shluková anlýza*. Praha, 2007.

[9] HONG, Yi, Sam KWONG, Yuchou CHANG a Qingsheng REN. Consensus Unsupervised Feature Ranking from Multiple Views. *Pattern Recognition Letters* [online]. 2008, vol. 29, issue 5, s. 595-602 [cit. 2015-05-18]. DOI: 10.1016/j.patrec.2007.11.012. ISSN 01678655.
Dostupné z: http://www.sciencedirect.com.

[10] BERKHIN, P. A Survey of Clustering Data Mining Techniques. In: KOGAN, Jacob, Charles K. NICHOLAS a Marc TEBOULLE. *Grouping Multidimensional Data: Recent Advance in Clustering*. Berlin: Springer-Verlag, 2006, s. 25-71. ISBN 978-3-540-28348-5.

[11] HAN, Jiawei a Micheline KAMBER. *Data Mining: Concepts and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006, 770 s. ISBN 15-586-0901-6.

[12] MAIMON, Oded a Lior ROKACH. *Data Mining and Knowledge Discovery Handbook*. 2nd ed. New York: Springer, 2010,1285 s. ISBN 978-0-387-09822-7.

[13] MURTAGH, Fionn, Pedro CONTRERAS. *Methods of Hierarchical Clustering* [online]. 2011. 21 s. [cit. 2015-05-18].
Dostupné z: http://arxiv.org/abs/1105.0121

[14] WANG, Li a Zheng-Ou WANG. CUBN: A Clustering Algorithm based on Density and Distance. In: *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics*. IEEE, 2003, s. 108-112. ISBN 0-7803-7865-2. DOI: 10.1109/ICMLC.2003.1264452.
Dostupné z: http://ieeexplore.ieee.org.

[15] GAN, Guojun, Chaoqun MA a Jianhong WU. *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia: SIAM, Society for Industrial and Applied Mathematics, 2007, 466 s. ISBN 978-0-898716-23-8.

[16] LIANG, Jiye, Xingwang ZHAO, Deyu LI, Fuyuan CAO a Chuangyin DANG. Determining the Number of Clusters Using Information Entropy for Mixed Data. In: *Pattern Recognition* [online]. 2012, vol. 45, issue 6, s. 2251-2265 [cit. 2015-05-18]. DOI: 10.1016/j.patcog.2011.12.017.
Dostupné z: http://www.sciencedirect.com.

[17] LIU, Huan a Lei YU. Toward Integrating Feature Selection Algorithms for Classification and Clustering. In: *IEEE Transactions on Knowledge and Data Engineering* [online]. 2005, vol. 17, issue 4, s. 491-502 [cit. 2015-05-18]. ISSN 10414347.
Dostupné z: http://ieeexplore.ieee.org.

[18] DY, Jennifer G., Carla E. Brodley. Feature Selection for Unsupervised Learning. In: *The Journal of Machine Learning Research* [online]. 2004, vol. 5, s. 845-889 [cit. 2015-05-18]. ISSN:1532-4435.
Dostupné z: http://dl.acm.org.

[19] YANG, Yang, Linxia LIAO, Guang MENG a Jay LEE. A Hybrid Feature Selection Scheme for Unsupervised Learning and its Application in Bearing Fault Diagnosis. In: *Expert Systems with Applications* [online]. 2011, vol. 38, issue 9, s. 11311-11320 [cit. 2015-05-18]. ISSN 09574174.
Dostupné z: http://www.sciencedirect.com.

[20] KRYSZCZUK, Krzysztof a Paul HURLEY. Estimation of the Number of Clusters Using Multiple Clustering Validity Indices. In: EL GAYAR, Neamat, Josef KITTLER a Fabio ROLI. *Multiple classifier systems: 9th International Workshop, MCS 2010, Cairo, Egypt, April 7-9, 2010: proceedings*. New York: Springer, s. 114-123. DOI: 10.1007/978-3-642-12127-2_12. ISBN 978-3-642-12127-2.

[21] DAVIES, David L. a Donald W. BOULDIN. A Cluster Separation Measure. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* [online]. 1979, PAMI-1, issue 2, s. 224-227 [cit. 2015-05-18].
DOI: 10.1109/TPAMI.1979.4766909.
Dostupné z: http://ieeexplore.ieee.org.

[22] ROUSSEEUW, Peter J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. In: *Journal of Computational and Applied*

Mathematics [online]. 1987, vol. 20, s. 53-65 [cit. 2015-05-18]. DOI: 10.1016/0377-0427(87)90125-7.
Dostupné z: http://www.sciencedirect.com.

[23] RENDÓN, Eréndira, Itzel Abundez, Alejandra Arizmendi a Elvia M. Quiroz. Internal Versus External Cluster Validation Indexes. In: *International Journal of Computers and Communications* [online]. 2011, vol. 5, issue 1, s. 27-34 [cit. 2015-18-05].
Dostupné z: http://w.naun.org/multimedia/UPress/cc/20-463.pdf.

[24] KOVACS, Ferenc, Csaba LEGÁNY a Attila BABOS. Cluster Validity Measurement Techniques. In: *6th International Symposium of Hungarian Researchers on Computational Intelligence* [online]. 2005, s. 18-29 [cit. 2015-05-18].
Dostupné z: http://citeseerx.ist.psu.edu/

[25] TIBSHIRANI, Robert, Guenther WALTHER a Trevor HASTIE. Estimating the number of clusters in a data set via the gap statistic. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* [online]. 2001, vol. 63, issue 2, s. 411-423 [cit. 2015-05-18]. DOI: 10.1111/1467-9868.00293.
Dostupné z: http://web.stanford.edu/~hastie/Papers/gap.pdf

[26] LIU, Huan a Rudy SETIONO. Scalable feature selection for large sized databases. In: *Proceedings of the 4th world conference on machine learning*. 1998, s 101-106.

[27] MOLINA, Luis Carlos, Lluís BELANCHE a Ángela NEBOT. Feature selection algorithms: a survey and experimental evaluation. In: *2002 IEEE International Conference on Data Mining. Proceedings* [online]. IEEE Comput. Soc, 2002, s. 306-313 [cit. 2015-05-18]. ISBN 0769517544.
DOI: 10.1109/ICDM.2002.1183917.
Dostupné z: http://ieeexplore.ieee.org.

[28] CAI, Deng, Chiyuan ZHANG a Xiaofei HE. Unsupervised feature selection for multi-cluster data. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10* [online]. New York, New York, USA: ACM Press, 2010, s. 333-342 [cit. 2015-05-18]. ISBN 9781450300551. DOI: 10.1145/1835804.1835848.
Dostupné z: http://dl.acm.org.

[29] CHANDRASHEKAR, Girish a Ferat SAHIN. A survey on feature selection methods. *Computers & Electrical Engineering* [online]. 2014, vol. 40, issue 1, s. 16-28 [cit. 2015-05-18]. DOI: 10.1016/j.compeleceng.2013.11.024.
Dostupné z:http://linkinghub.elsevier.com/retrieve/pii/S0045790613003066

[30] MITRA, P., C.A. MURTHY a S.K. PAL. Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine*

Intelligence [online]. Vol. 24, issue 3, s. 301-312 [cit. 2015-05-18]. DOI: 10.1109/34.990133.
Dostupné z: http://ieeexplore.ieee.org.

[31] MASAELI, Mahdokht, Yan YAN, Ying CUI, Glenn FUNG a Jennifer G. DY Convex Principal Feature Selection. In: *Proceedings of the 2010 SIAM International Conference on Data Mining* [online]. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2010, s. 619-628 [cit. 2015-05-18]. ISBN 9780898717037.
Dostupné z: http://citeseerx.ist.psu.edu.

[32] YANG, Yi, et al. l2, 1-norm regularized discriminative feature selection for unsupervised learning. In: *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* [online]. 2011, vol. 22, no. 1, s. 1589-1594 [cit. 2015-05-18]. ISBN: 978-1-57735-514-4. DOI: 10.5591/978-1-57735-516-8/IJCAI11-267.
Dostupné z: http://ijcai.org/papers11/Papers/IJCAI11-267.pdf.

[33] QINPEI Zhao. Cluster Validity in Clustering Methods. *Accademic Dissertation in Forestry and Natural Sciences No 77*. 2012.

[34] NATTHAKAN, Iam-on a Simon Garrett. LinkCluE: A MATLAB Package for Link-Based Cluster Ensembles. In: *Journal of Statistical Software* [online]. 2010, vol. 36, issue 9, s. 1-36 [cit. 2015-05-18]. ISSN: 1548-7660.
Dostupné z: http://www.jstatsoft.org/v36/i09

[35] BLAKE, Cathy a Christopher MERZ. UCI Machine Learning Repository. Irvine, CA: University of California, Schol of Information and Computer Science. 1998.
Dostupné z: http://archive.ics.uci.edu/ml.

[36] LITTLE, Max A., Patrick E. MCSHARRY, Stephen J. ROBERTS, Declan AE. COSTELLO a Irene M. MOROZ. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. In: *BioMedical Engineering* [online]. 2007, vol 6, issue 23 [cit. 2015-05-18]. DOI: 10.1186/1475-925X-6-23.
Dostupné z: http://www.biomedical-engineering-online.com/content/6/1/23.

[37] TSANAS, Athanasios, Max A. LITTLE, Cynthia FOX a Lorraine O. RAMIG. Objective Automatic Assessment of Rehabilitative Speech Treatment in Parkinson's Disease. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* [online]. 2014, vol. 22, issue 1, s. 181-190 [cit. 2015-05-18]. DOI: 10.1109/TNSRE.2013.2293575. ISSN 15344320.
Dostupné z: http://ieeexplore.ieee.org.

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM GRAFŮ

SEZNAM PŘÍLOH

Příloha 1: CD obsahuje:

- Diplomová práce ve formátu PDF
- Zdrojové kódy k programu pro experimentální ověření metod shlukové analýzy